Probing Protein Structural Dynamics using Simplified Models

Yiwen Chen

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Physics and Astronomy

Chapel Hill
2007

Approved by:

Dr. Richard Superfine
Dr. Nikolay V. Dokholyan
Dr. Sharon L. Campbell
Dr. Sean Washburn
Dr. Yue Wu

©2007
Yiwen Chen
ALL RIGHTS RESERVED

ABSTRACT

YIWEN CHEN: Probing Protein Structural Dynamics using Simplified Models
(Under the direction of Dr. Nikolay V. Dokholyan)

Structure and dynamics play crucial roles in many aspects of protein function. The detailed characterization of structure and dynamics of a protein is therefore critical for elucidation of its function. Despite the rapid advances in experimental methods, we are still much limited in our ability to characterize the structure and dynamics of a protein due to the resolution capability at length and time scales of the current methods. As complimentary approaches to the experimental methods, coarse-grained simulations based on simplified models offer unparallel opportunities for studying structure and dynamics that are hardly subject to direct experimental characterization.

In this dissertation, we first develop a new methodology that enables the study of the meta-stable states of protein by incorporating hydrogen-exchange data derived from NMR experiments into coarse-grained simulations, and applied this method to characterize the folding intermediate of FAT domain. Second, we study large-scale conformational dynamics of vinculin and reveal new insights into allosteric control of its function. Third, we study the

fidelity of protein structure reconstruction using inter-residue proximity constraints and rational strategies for constraint selection for protein structure determination.

Dedicated to my parents,

and my wife Jian.

ACKNOWLEDGEMENTS

I would like to thank many people for their help along the way of getting my PhD. I am mostly grateful to my advisor Professor Nikolay V. Dokholyan, for his supervision and guidance in all research projects. I would like to thank Professor Sharon L. Campbell who collaborated with us on the project of FAT domain (Chapter 1) and other works that are not included in this dissertation, for her great help in many aspects of my research. I would also like to thank other experimental collaborators in the FAT domain project: Professor Michael D. Schaller, Dr. Richard Dixon and Dr. Kirk C. Prutzman from Campbell's lab. In addition, I would like to thank my committee members Professor Richard Superfine, Sean Washburn and Yue Wu, who have given me valuable suggestions on my research works. Last but not the least; I would like to thank Dr. Feng Ding, Dr. Sagar Khare, Kyle Wilcox and other members from Dokholyan's lab.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

I. INTRODUCTION

General background 1

Experimental tools for studying protein structure and dynamics 7

Theoretical/computational methods 16

II. COMBINING HYDROGEN EXCHANGE DATA AND COARSE-GRAINED SIMULATION TO CHARACTERIZE FOLDING INTERMEDIATE OF FAT DOMAIN

Introduction 23

Materials and methods 25

Results and discussion 29

III. CHARACTERIZING THE LARGE-SCALE STRUCTURAL DYNAMICS OF VINCULIN RELATED TO ITS ACTIVATION

Introduction 42

Materials and methods 44

Results and discussion 48

IV. DETERMINING PROTEIN STRUCTURE USING INTER-RESIDUE PROXIMITY CONSTRAINTS

Introduction 63

Materials and Methods 65

Results and Discussion Conclusions 69

Conclusions 80

REFERENCES 81

LIST OF TABLES

Table

4.1 Nine protein domains 70

4.2 Two large protein domains 74

4.3 The differences in the edge betweenness centrality between the constraint sets 78

LIST OF FIGURES

Figure

1.1 The reaction scheme of hydrogen exchange 15

1.2 An illustration of the principle of hydrogen exchange experiments 16

1.3 The two principal components of a sample two-dimensional data set 17

1.4 A flowchart illustration of Monte Carlo Metroplis algorithm 19

1.5 Difference between traditional molecular dynamics and discrete molecular dynamics simulation 20

1.6 A sample graph and two sub-graphs 22

2.1 Energy map for the FAT domain 30

2.2 Correlation between experimental and calculated protection factors 31

2.3 The folding trajectory for the scaled DMD simulations 33

2.4 The probability distribution at three simulation temperatures 34

2.5 Frequency map for the intermediate ensemble 35

2.6 Comparison of the FAT domain intermediate from DMD simulations with the domain-swapped dimer 37

3.1 The native structure of full-length vinculin 50

3.2 The thermal melting curves of inter-domain and intra-domain interactions 52

3.3 A cartoon representation of the sequential unfolding events in the kinetic simulation 54

3.4 A sample trajectory from kinetic unfolding simulations 55

3.5 The fraction of native contacts $Q_{D1\text{-}Vt}$ made between D1 and Vt domains is plotted as a function of temperature 56

4.1 The average and the standard deviation of RMSD of the structural ensembles for nine small protein domains 72

4.2 The average and the standard deviation of RMSD of the structural ensembles for two large protein domains 73

4.3 The average and standard deviation of RMSD of the structural ensembles for Neurotoxin b (PDB code: 1NXB) 76

4.4 The histograms of contact distance for Neurotoxin b (PDB code: 1NXB) 77

ABBREVIATIONS

ECM	Extracellular matrix
FAK	Focal adhesion kinase
PTM	Posttranslational modification
NMR	Nuclear magnetic resonance
HSQC	Heteronuclear single quantum correlation
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect Spectroscopy
RF	Radio frequency
ESI	Electrospray ionization
MALDI	Matrix-assisted laser desorption/ionization
MS/MS	Tandem mass spectrometry
HX	Hydrogen exchange
PCA	Principal component analysis
DMD	Discrete molecular dynamics
FAT domain	Focal adhesion targeting domain
Vt	Vinculin tail domain
VASP	Vasodilator-stimulated phosphoprotein
Arp2/3	Actin-related protein complex
RMSD	Root-mean-square-deviation
FRET	Fluorescence resonance energy transfer
EPR	Electron paramagnetic resonance
PRE	Paramagnetic relaxation enhancement

MS	Mass spectrometry
FGF	Fibroblast growth factor
COR	Contact order-ranked

CITATIONS TO PUBLISHED WORKS

Chapter 2, 3, and 4 were reproduced with modifications based on three published works. The permissions for the reproduction of figures and texts in the dissertation were granted by the publishers.

Chapter 2 was published in *Structure 12*, Dixon, R.D., Chen, Y., Ding, F., Khare, S.D., Prutzman, K.C., Schaller, M.D., Campbell, S.L., and Dokholyan, N.V. (2004), *New insights into FAK signaling and localization based on detection of a FAT domain folding intermediate*, 2161-2171.
Copyright © 2004 Elsevier Ltd.

Chapter 3 was published in *Journal of Biological Chemistry 281,* Chen, Y. and Dokholyan, N.V. (2006), *Insights into allosteric control of vinculin function from its large-scale conformational dynamics*, 29148-29154.
Copyright © 2006 the American Society for Biochemistry and Molecular Biology

Chapter 4 was reproduced in part with permission from *Journal of Physical Chemistry B 111*, Chen, Y., Ding, F., and Dokholyan, N.V. (2007), *Fidelity of the protein structure reconstruction from inter-residue proximity constraints*, 7432-7438.
Copyright © 2007 American Chemical Society

CHAPTER 1

INTRODUCTION

General background

Protein

Proteins are linear biopolymers that are composed of 20 different amino acids(Branden and Tooze, 1999). All amino acids share a common chemical structure including a central carbon (C_α) atom to which an amino group (NH_2), a carboxyl (C'OOH) and a side chain are attached. The major difference that distinguishes one type of amino acid from another is its side chain. The amino acids in a protein are covalent-linked together by peptide bonds (C'-N) which are formed between the amino group of one amino acid and the carboxyl group of the next through dehydration reactions. The peptide bond has some double bond characteristics and therefore it is unable to rotate freely around it, while rotation is permitted about the N-C_α and C_α-C' bonds. The main-chain or backbone atoms in a protein refer to C_α atom, an NH group bound to C_α atom, and a carbonyl group C'=O, where C' is attached to C_α. The two dihedral angles around N-C_α and C_α-C' bonds determine the local geometries assumed by the protein backbone. The unique chemical structure of amino acids results in the directionality of a protein: one end of the protein with a free amino group is known as amino terminus or in short N-terminus, while the other end of the protein with a free carboxyl group is known as carboxy terminus or C-terminus.

Primary, secondary, tertiary and quaternary structure

Most proteins fold into a unique three-dimensional structure which is known as its native structure/state(Branden and Tooze, 1999). A three-dimensional structural unit that can fold independently within a protein is called a domain. A protein is known as single-domain if it has only one domain, otherwise it is called a multi-domain protein. A series of terminologies are used to describe different hierarchies of protein structure. Primary structure refers to the ordered sequence of amino acids (from N-terminus to C-terminus) that compose the protein. Secondary structure refers to the regularly repeating local structures that are stabilized by hydrogen bonds. The most common examples of secondary structure are the alpha helix and beta sheet. The spatial arrangement of the secondary structures i.e. overall topology of a single protein is known as its tertiary structure. The overall three-dimensional structure of a protein complex that results from the assembly of more than one protein is known as quaternary structure. Each protein in the protein complex is usually called protein subunits. Proteins can be roughly classified as globular proteins, fibrous proteins, and membrane proteins. Most of the globular proteins are soluble; while many membrane proteins are insoluble in water and therefore more difficult to study. Fibrous proteins often play structural roles in cellular functions.

It is believed that proteins can carry out their biological functions only when they adopt correct native structure, because three-dimensional shape of the proteins in the native state is essential to their function.

Extracellular matrix

Extracellular matrix (ECM) is the part of a tissue that is not part of any cell in the tissue(Lodish et al., 2004). The main components of the ECM are proteoglycans , glycoproteins, and hyaluronic acid. It also contains proteins including fibrin, elastin, fibronectins, lamins, and nidogens, and other substances such as minerals or fluids. Moreover, it acts as a reservoir for a wide variety of growth factors, the release of which can cause the rapid alteration of physiological states of cells. In overall, it serves as a micro-enviroment for the cells by providing support and anchorage for cells as well as regulating cell signaling. It is essential to maintaining the physiological structure and function of a tissue.

Focal adhesion

Focal adhesions are integrin-rich cell adhesion sites that make close contact to the ECM(Lodish et al., 2004). This physical interaction confers cells the ability to communicate with outside environment, which is essential for cell adhesion, migration, proliferation and death. At molecular level, focal adhesions are comprised of large molecular complexes forming around an integrin heterodimer. Integrin binds to ECM proteins through its extracellular domain. The cytoplasmic domain of integrin binds to the cytoskeleton through adapter proteins such as α-actinin, filamin, talin and vinculin. Many cell signalling proteins, such as focal adhesion kinase (FAK), associate with this integrin-adapter protein-cytoskeleton complex, which is the molecular basis for a focal adhesion to connect the actin cytoskeletons to the cytoplasmic side of the membrane, and mediates intracellular signaling(Lo, 2006). Focal adhesions not only serve for the anchorage of the cell to the ECM, but also can function as major ‘sensors’ of various biochemical signals and mechanical

stimuli, which inform the cell about the micro-environment. In non-moving cells, focal adhesions are relatively stable, whereas in moving cells they constantly assemble and disassemble as the new contacts form at the leading edge, and dissemble at the trailing edge of the cell when old contacts are broken.

Cell-cell adhesion

The adhesion of one cell to another is critical to the formation and maintenance of tissues in multi-cellular organism(Lodish et al., 2004). At cell-cell contact sites, the extracellular domains of transmembrane adhesion molecules interact with molecules on the surface of adjacent cells, and the cytoplasmic domains are associated with the cytoskeleton through various adaptor proteins. Connecting the cytoskeleton to adjacent cells confers mechanical strength to tissues and provides physical basis for the cytoskeletal movements that mediate changes of cell morphologies during development. Cell-cell adhesion is tightly controlled during processes that require detachment and reattachment between cells, such as cell proliferation and morphogenesis. In cancers of solid tissues, one of the hallmarks that feature the transition of to a malignant, invasive tumor is the loss of regulated cell adhesion(Hanahan and Weinberg, 2000). Cell-cell adhesion sites vary significantly at molecular level, but they share a general architecture in which a transmembrane protein is linked to the cytoskeleton via adaptor proteins. Different cell–cell adhesion sites were originally defined as distinct structures by electron microscopy based on their macroscopic morphologies, including tight junctions, adherens junctions and desmosomes(Lodish et al., 2004). Tight junctions and adherens junctions are connected to the actin-based cytoskeleton. In contrast, desmosomes

are connected to the intermediate filaments. In all of these junctions, the presence of adaptor proteins in the adhesion molecule–cytoskeletal linkage provides targets for regulatory signals that control the strength and assembly of cell contact sites. In general, cell adhesion proteins are classified based on the structure of the adhesion proteins and their corresponding ligands. Adhesion between two molecules of the same adhesion protein is called "homophilic" binding, and adhesion between an adhesion protein and a different type of protein is called "heterophilic" binding(Lodish et al., 2004).

Cell migration

Cell migration is a central process in the physiology of multi-cellular organisms. It plays essential roles in the processes such as embryonic development, wound healing and immune responses. The aberrant migration of cells is often related to serious human diseases including mental retardation, vascular disease, rheumatoid arthritis, and cancer. Therefore the understanding the basic mechanisms of cell migration holds the promise of effective therapies for treating diseases(Ridley et al., 2003).

In general, cell migration can be considered as a cyclic process(Lauffenburger and Horwitz, 1996; Ridley et al., 2003). The initial stage of migration is the polarization and extension of protrusions of cells in response to migration-promoting factors in the direction of migration. These protrusions are usually driven by actin polymerization, and are stabilized by forming adhesions to the ECM or adhering to adjacent cells via transmembrane receptors connected to the actin cytoskeleton. These adhesions serve as anchoring point or traction sites for cell migration, and they are disassembled at the rear of the cell, which allow cells to move forward. Formation and disassembly of adhesions occur in a dynamic fashion, the rate

of which is often correlated with the motility of a cell(Friedl and Wolf, 2003; Knight et al., 2000; Ridley et al., 2003).

Although, this general picture is shared by many cell types, the details may differ significantly. For example, this cyclic process is pronounced in slow-moving cells such as fibroblast, but not as obvious as in fast-moving cells such as neutrophils. In addition, the way how cell migrates highly depend on its environment. For instance, somitic cells migrating in vivo exhibit large single protrusions and highly directed migration, which is distinct from the multiple small protrusions they show on planar substrates; cancer cells are able to modify their migratory behavior and morphology in response to environmental changes(Knight et al., 2000; Ridley et al., 2003).

As a highly coordinated process, cell migration involves a complex network of interaction between different proteins. FAK and vinculin, which are two major subjects of study in this dissertation, play significant roles in this process.

Phosphorylation

After protein translation, the posttranslational modification (PTM) of amino acids can alter protein function by attaching to it other biochemical functional groups such as acetate, phosphate, various lipids and carbohydrates, which changes the chemical nature of an amino acid. Phosphorylation is the addition of a phosphate (PO_4) group to a protein and an important form of PTM(Lodish et al., 2004). It can occur on serine, threonine, tyrosine, histidine and aspirate. Phosphorylations on histidine and aspartate occur in prokaryotes as a major mode of signal transduction in two-component signaling. In eukaryotes, protein phosphorylation is one of the major regulatory events at post-translational level. Many

proteins are activated or deactivated by phosphorylation and dephosphorylation. Phosphorylation is catalyzed by various specific protein kinases, whereas dephosphorylation of a protein is catalyzed by the machineries, called phosphatases.

Experimental methods for studying protein structure and dynamics

Nuclear magnetic resonance spectroscopy

Protons and neutrons have a spin angular momentum with a value of +1/2 or -1/2. Protons or neutrons in the atomic nucleus can pair with other protons or neutrons with anti-parallel spin angular momentums. A nucleus with all protons and neutrons that are in pair has a net spin angular momentum of zero, but a nucleus with unpaired protons or neutrons will have a non-zero overall spin. When the spin angular momentum of a nucleus is non-zero, it has an associated magnetic moment μ, which is utilized for manipulation in nuclear magnetic resonance (NMR) experiments(Cavanagh et al., 1996). The nuclei that have odd numbers of nucleons (protons and neutrons) have a non-zero intrinsic magnetic moment. The most commonly used nuclei in NMR experiments are 1H, 13C and 15N. NMR studies the nuclear spin dynamics of a magnetic nucleus by aligning its nuclear spin with a large external magnetic field and perturbing this alignment using an electromagnetic pulse field. The response of nuclei to the perturbation using pulse field is what is monitored in NMR spectroscopy. The response reflects the environment of nuclear spins(Cavanagh et al., 1996). NMR spectroscopy is one of the principal techniques used to obtain structural and dynamic information of a protein. It is a powerful technique that can provide detailed information on the three-dimensional structure and dynamic motions of proteins in solution.

Nuclear spin angular momentum is a vector quantity. The Z component (the component along the direction of external magnetic field) of the nuclear spin angular momentum, I_z can take values ranging from $+I$ to $-I$ in integer steps where I is the magnitude of nuclear spin angular momentum(Cavanagh et al., 1996). For a given nucleus with spin angular momentum I, there are in total $(2I+1)$ angular momentum states along Z-axis. Therefore the Z component of spin angular momentum I_z, is quantized as follows:

$$I_Z = m\frac{h}{2\pi}(-I \le m \le I)$$

where h is Planck's constant and π is a mathematical constant, i.e. the ratio of a circle's circumference to its diameter in Euclidean geometry . The magnetic moment of this nucleus is related to its spin angular momentum with a proportionality constant γ, which is called the gyro-magnetic ratio:

$$\mu_Z = \gamma I_Z$$

For a nucleus that has a spin of one-half (I=1/2), the nucleus has two possible Z component of the spin angular momentum I_z: +1/2 or -1/2 (up or down state). The energies of these two spin states are degenerate in the absence of external magnetic field. When a nucleus is subject to a magnetic field, the two magnetic momentum states no longer have equal energy since the energy of a magnetic moment μ in a magnetic field B_0 is the negative scalar product of the two vectors:

$$E = -\mu_Z B_0 ,$$

where μ_z is the Z component of the nuclear magnetic moment and magnetic field is along Z-axis. We can substitute the relationship of, $\mu_z=\gamma I_z$ to the above equation and obtain:

$$E = -\frac{mh\gamma B_0}{2\pi}(m = -\frac{1}{2}, \frac{1}{2}),$$

The energy gap between the two spin states is then $(h\gamma B_0)/2\pi$. A resonance can occur between these two states when a radiofrequency (RF) is applied with the same energy as the energy difference ΔE between the two spin states. The energy of a RF photon is $E = hv$, where v is its frequency.

$$\Delta E = \frac{h\gamma B_0}{2\pi}$$

Thus, the frequency of electromagnetic radiation required to produce resonance of a specific type nucleus in a field B_0 is:

$$v_0 = \frac{\gamma B_0}{2\pi}$$

This resonance gives rise to the nuclear magnetic resonance spectrum. Since other nuclei, especially spin-active nuclei, and local electron are able to shield each probed nucleus differently from the main external field B_0, the strength of the effective magnetic field $B_{effective}$ at the nucleus is different from the applied magnetic field B_0. Therefore the frequency necessary to achieve resonance becomes $\frac{\gamma B_{eff}}{2\pi}$ and is different from the expected value of $\frac{\gamma B_0}{2\pi}$. The relative deviation of resonance frequency with respect to certain reference

frequency $\frac{v_{eff} - v_{ref}}{v_0}$ is called chemical shift. For nuclei in different chemical environments, the chemical shifts are usually different(Cavanagh et al., 1996). The chemical shit differences between nuclei give rise to distinct peak frequencies in a nuclear magnetic resonance spectrum, which forms the basis for NMR spectroscopy to be a direct probe of chemical environment of a nucleus. The chemical environment of a nucleus in a protein often changes when a protein undergoes transitions between several conformations, which are induced by the binding of a substrate to the active site of an enzyme, or a protein ligand to the protein(Cavanagh et al., 1996).

Heteronuclear Single Quantum Correlation

A heteronuclear single quantum correlation (HSQC) is an experiment frequently used in protein NMR spectroscopy(Cavanagh et al., 1996). The HSQC spectrum has two axes, a proton axis and a hetero-nuclei axis. A hetero-nucleus is another nucleus other than protons, which is most often 13C or 15N. The peaks in the HSQC spectrum usually correspond to different protons attached to hetero-nuclei.

The 15N HSQC experiment is one of the most frequently applied experiments in protein NMR, which is often performed using isotopically-labeled protein. The amide protons attached to nitrogens in the peptide bonds of most residues in a protein (except proline) give rise to peaks in the HSQC spectrum. If the studied protein is well folded, the peaks are usually well dispersed and therefore can be distinguished due to the distinct chemical environment of different amide protons. The assignment of the 15N spectrum is usually

essential for interpretation of more advanced NMR experiments and an important procedure for protein structure determination.

The HSQC experiment is also useful for detecting protein-protein, protein-ligand or protein-drug interactions. By comparing the HSQC of the free state of a protein with its bound-state, it is possible to find out the binding interface where the chemical shifts of the peaks are most likely to change.

Nuclear Overhauser Effect Spectroscopy

Overhauser Effect generally refers to the transfer of spin polarization from one spin population to another. Overhauser effect can occur between electrons or atomic nuclei. The spin polarization transfer is commonly observed and used amongst atomic nuclei and named Nuclear Overhauser Effect (NOE)(Cavanagh et al., 1996). A commonly-applied technique in structural biology that is based on NOE is Nuclear Overhauser Effect Spectroscopy (NOESY). NOESY spectra provide distance information about protons that are within 5 Angstroms. The NOESY spectrum obtained in the experiment has two proton axes. The presence of a NOE peak between two protons indicates that the corresponding protons are within 5 Angstroms through space and the intensity of the peak is proportional to $\frac{1}{r^6}$, where r is the distance between two protons(Cavanagh et al., 1996). It should be noted that the absence of a NOE peak between two protons does not necessarily mean that they are not within 5 Angstroms since there are other factors that can reduce the intensity of a NOE peak.

Mass spectrometry

Mass spectrometry is a powerful tool that is used to analyze and identify the composition in a mixture based on the difference of the mass-to-charge ratio of constituent components. The basic idea of mass spectrometry is that by applying electric and magnetic fields to the ionized molecules, different components in the samples have different modes of motions in space or time, due to their difference in mass-to-charge ratio, so that one can separate and detect them based upon this spatial or temporal difference. Due to its incomparable power, it is becoming an indispensable tool in identifying specific proteins and the post-translational modifications of the proteins in a complex biological sample. There are three major components that constitute a mass spectrometry: ion source, mass analyzer, and detector.

The ion source ionizes the material of interest and the generated ions are then passed to the mass analyzer by applying magnetic or electric fields. Ionization techniques have been key determinants for the type of samples (liquid or solid samples) to be analyzed by mass spectrometry. There are two major techniques that are used with liquid and solid biological samples, respectively: electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI)(Hoffmann and Stroobant, 2001).

ESI is a technique that generates ionized liquid droplets through electrostatic charging. In ESI, liquid sample is first passed through a nozzle. The droplets are then generated by electrically charging the liquid in the nozzle by applying a very high voltage. The liquid in the nozzle becomes unstable when it is forced to carry more and more charge. Once a critical point is reached, at which the liquid is unable to hold more electrical charge, it blows apart into tiny and highly charged droplets at the tip of the nozzle. ESI is the primary ion source

used in liquid chromatography-mass spectrometry because it provides a natural liquid-gas interface that is capable of coupling liquid chomatography with mass spectrometry(Hoffmann and Stroobant, 2001).

In contrast, in MALDI, the ionization is triggered by a laser beam shedding on a solid matrix that is used to protect proteins from being damaged by direct shedding of laser beam(Hoffmann and Stroobant, 2001). In MALDI, a chemical solvent is mixed with the protein molecule of interest and then spotted onto a MALDI plate. The solvent molecules vaporize, leaving the proteins and the matrix co-crystallized in a MALDI spot. When a laser is shed at the crystals in the MALDI spot, the spot absorbs the laser energy and the matrix becomes ionized. The matrix transfers part of its charge to the protein molecule, thus ionizing them while still protecting them from the possible destruction of the laser.

The mass analyzer is used to separate molecules by applying electric and magnetic fields to change the motion modes of ions in space and time(Hoffmann and Stroobant, 2001). Mass analyzers separate the ions according to their mass-to-charge ratio based upon the dynamics of charged particles in electric and magnetic fields in vacuum, which is described by the following equation:

$$(\frac{m}{q})\vec{a} = \vec{E} + \vec{v} \times \vec{B} ,$$

where *m* is the mass of the ion, $\vec{a}$ is the acceleration, *q* is the ionic charge, $\vec{E}$ is the electric field, and $\vec{v} \times \vec{B}$ is the vector cross product of the ion velocity and the magnetic field

This differential equation describes the classic equation of motion of charged particles. It completely determines the particle's motion in space and time when the particle's initial conditions are defined, and therefore forms the physical basis of every mass analyzer. It follows from this equation that two particles with the same physical quantity *m/q* behave exactly the same. Thus all mass spectrometry measure *m/q*. Different mass analyzers use either static or dynamic fields, or magnetic or electric fields, but all operate based on the same equation. The most commonly-used analyzers are time-of-flight, quadrupole and quadrupole ion trap. Many mass spectrometry use two or more mass analyzers for tandem mass spectrometry (MS/MS)(Hoffmann and Stroobant, 2001).

The last but not the least component of the mass spectrometry is the detector. When an ion passes by or hits a surface, the charge induced or current produced is recorded by the detector(Hoffmann and Stroobant, 2001). The recorded signal then gives rise to a mass spectrum comprised of a record of ions as a function of *m/q*.

Hydrogen Exchange (HX)

Hydrogen exchange (HX) is a chemical reaction in which a covalently bonded proton in a protein is replaced by a deuteron from solvent. Usually the examined protons are the amide proton in the backbone(Englander et al., 1996; Englander et al., 1997). HX of amide proton can be monitored by either NMR spectroscopy or mass spectrometry. For a typical NMR HX experiment, a protein of interest is put into deuterium water D_2O. HSQC spectra are then recorded at a series of time points while the amide hydrogen is exchanging with the deuterium from D_2O. The signal of amide proton will decay exponentially as the proton

exchanges. The exchange constant is then obtained by fitting an exponential function to the data. When an amide proton is buried in a protein or forms intermolecular hydrogen bonds, it is protected against exchange with deuterons from solvent(Krishna et al., 2004). A protein has to undergo a conformational change from a well-folded state (C) in which the given amide proton is protected from exchange to an open state (O) where the same amide proton is competent for exchange (Figure 1.2)(Englander et al., 1996). Such a reaction scheme is shown in Figure 1.1., where k_{open} is the first-order rate constant for the transition from the folded protein to the open state and k_{close} is the rate constant of the reverse transition. k_{rc} is the intrinsic rate constant for exchange of amide proton with deuterium atom. Under EX2 condition where $k_{close} >> k_{rc}$, the effective exchange rate constant is $k_{ex} = \frac{k_{open}}{k_{close}} k_{rc} = K_{op} k_{rc}$ (Clarke and Fersht, 1996; Englander, 2000).

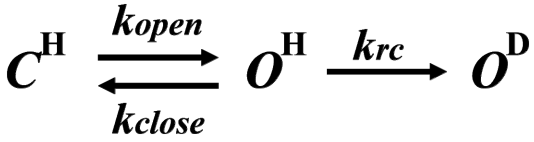

FIGURE 1.1. The reaction scheme of hydrogen exchange

Therefore by dividing the effective exchange rate constant k_{ex} with tabulated intrinsic rate constant k_{rc} for exchange of amide proton, the equilibrium constant of opening reaction K_{op} can be calculated and the free energy corresponding to the opening reaction can be obtained(Clarke and Fersht, 1996; Englander, 2000).

Given its capability in probing the structural information of weakly-populated states of a protein that are difficult to study by other experimental methods, HX experiment is an important tool for elucidating protein folding pathways, protein dynamics and protein-protein interactions(Clarke and Fersht, 1996; Englander, 2000; Wand and Englander, 1996).

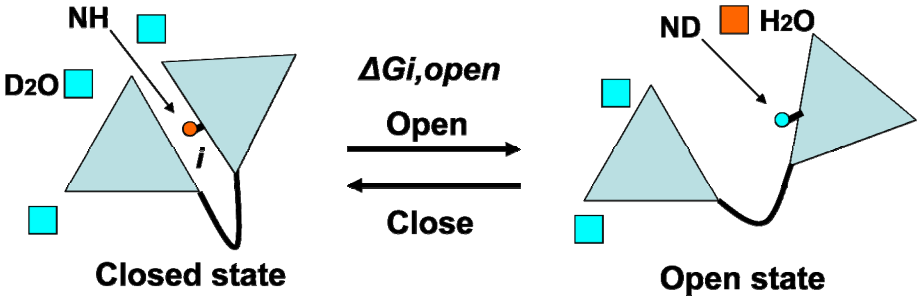

FIGURE 1.2. An illustration of the principle of hydrogen exchange experiments. A protein has to undergo a conformational change from a closed state in which the given amide proton is protected, to an open state where the same amide proton is competent for exchange. Under EX2 limit, the free energy cost of opening reaction of a given amide proton can be estimated from HX experiment.

Theoretical/computational methods

Principle Component Analysis

Principal components analysis (PCA) is a statistical technique for simplifying the representation of a data set, by reducing multidimensional data sets to lower dimensions(Jolliffe, 2002). Essentially, it is an orthogonal linear transformation that transforms the data to new coordinate so that the greatest variance by any projection of the

data lies on the first coordinate axis in the new coordinates, the second greatest variance on the second coordinate axis, and so on. The coordinate axes along which the greater variances of the data lie are called principal components. By keeping lower-order principal components and ignoring higher-order ones, PCA are used to reduce the dimensionality of the data set while retaining those features of the data set that contribute most to its variance. For example, in a two-dimensional data set that is shown in Figure 1.3, the two principle component axes are shown in orange. It is clear from figure that the variations of this data set along the first principal component accounts for most of variations in this dataset.

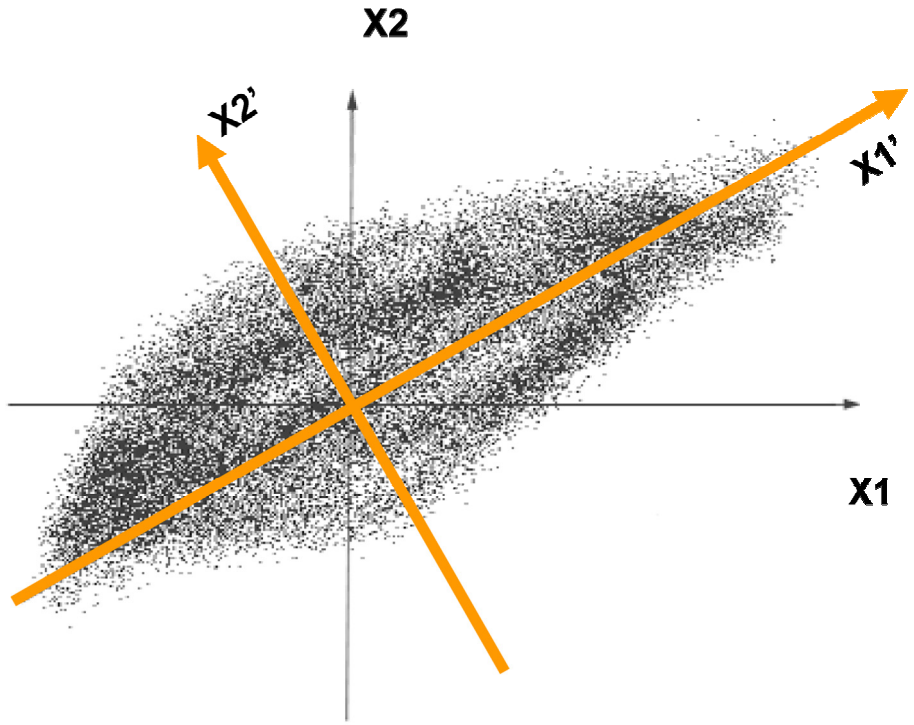

FIGURE 1.3. The two principal components of a sample two-dimensional data set are shown in orange.

Monte Carlo Metroplis algorithm

In this algorithm(Binder, 1995; Liu, 2001), a stochastic search is performed in the configuration space of interest through multiple iterations. The stochastic search is designed so that the points visited in the configuration space are distributed according to certain probability distribution that is of interest. At each point on the path of stochastic search, a random trial move from the current position in configuration space is generated. This trial move is then either rejected or accepted according to certain probabilistic rule. If the move is accepted then the search proceeds starting from the new position in configuration space; otherwise there is no move in configuration space. Another trial move is then generated, either from the newly accepted position or from the old position if the first trial was rejected, and the process is repeated until the stochastic search has converged in the configuration space. The Metroplis algorithm(Liu, 2001) is illustrated in Figure 1.4, where E' represents the coordinate of the current position in configuration space (assuming the space is one-dimensional) and ΔE represents the random trial move; W is the cost function and ΔW is the change of cost function when the position is changed.

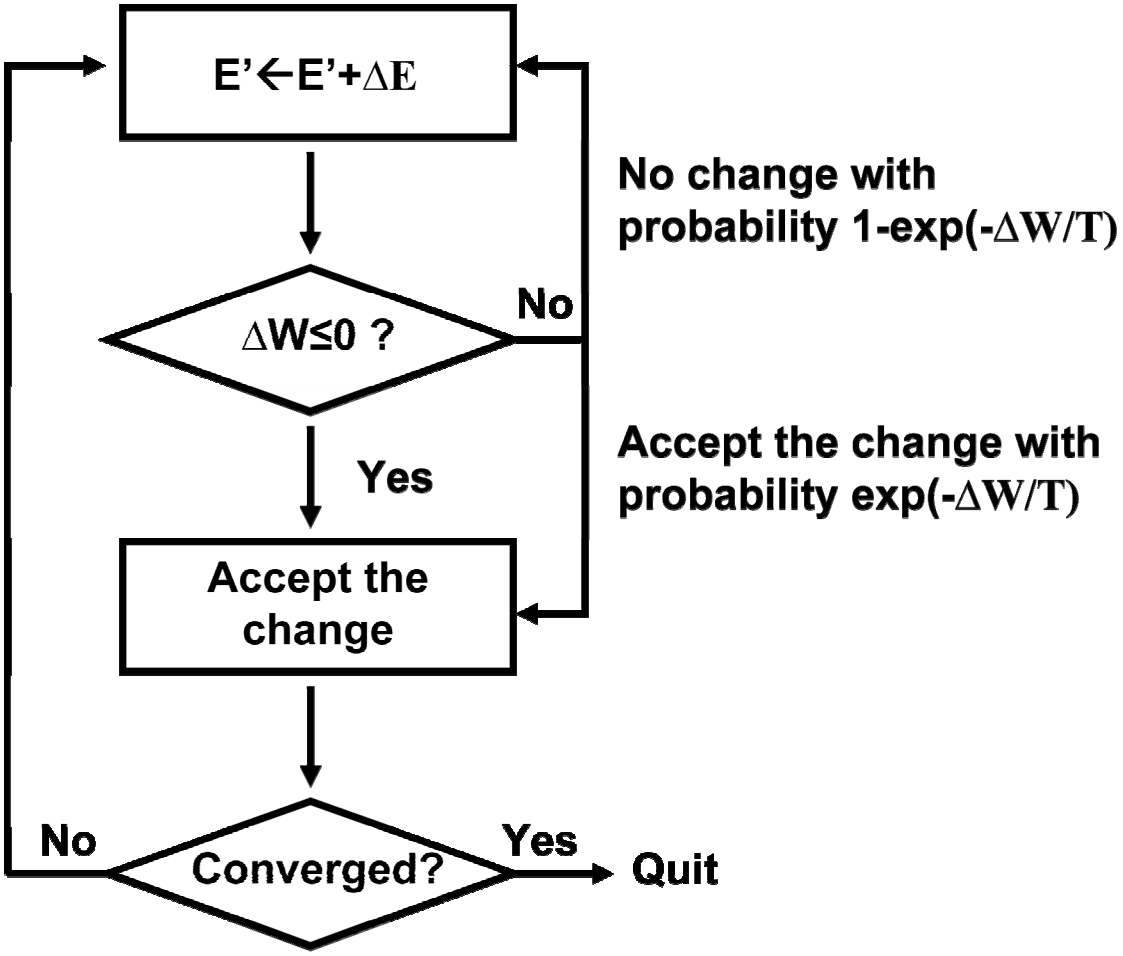

FIGURE 1.4. A flowchart illustration of Monte Carlo Metroplis algorithm

Molecular dynamics

Molecular dynamics is a general simulation method that is used to study the thermodynamic and kinetic properties of a given system by solving the Newtonian equation of the system(Rapaport, 2004). There are different molecular dynamics simulations that have different computational efficiencies and resolution. In all-atom molecular dynamics simulation, each atom in a protein is modeled and therefore high-resolution structural

information about protein motion can be provided. In contrast, in coarse-grained molecular dynamics simulation, only a small fraction of representative atoms in a protein are modeled and has higher computational efficiencies compared with all-atom simulations

Discrete Molecular Dynamics

In general, discrete molecular dynamics (DMD) simulation(Dokholyan et al., 1998; Rapaport, 2004; Smith et al., 1997) is based on pair-wise spherically symmetrical potentials that are discontinuous step-well functions of inter-particle distances (Figure 1.5). In DMD, all particles move with a constant velocity unless they cross the boundaries of the step-well potentials. At the moment of crossing boundaries, their velocities change instantaneously. This change satisfies the laws of energy, momentum, and angular momentum conservation. Each time, the next soonest boundary-crossing is determined and the state of the system is updated to the time point when this boundary-crossing occurs. Therefore DMD simulation saves the calculation of the particle motions between potential boundaries.

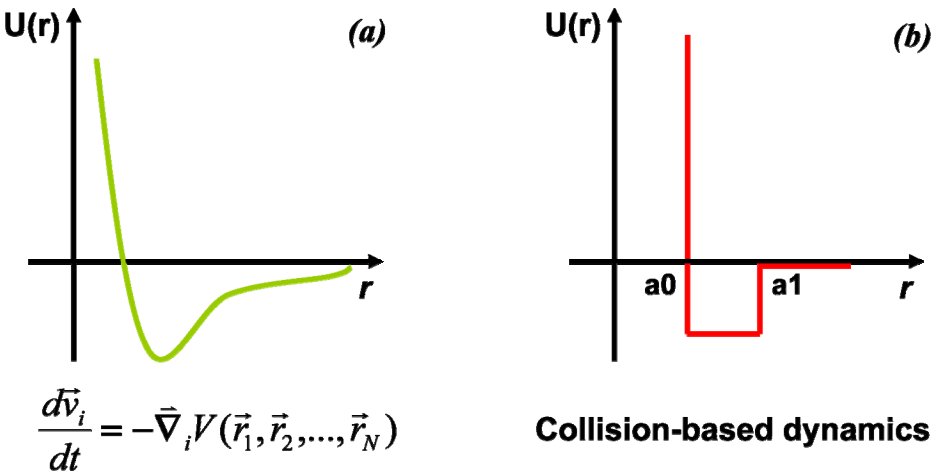

FIGURE 1.5. Different molecular dynamics simulation methods: *(a)* traditional molecular dynamics and *(b)* discrete molecular dynamics simulation are shown.

Graph theory

Graphs are mathematical structures used to model pair-wise relations between objects(Diestel, 2005; Albert and Barabasi, 2002). A graph is composed of a collection of nodes and a set of edges that connect pairs of nodes (Figure 1.6). A graph is called undirected if there is no distinction between the two nodes associated with each edge; or directed if its edges are directed from one node to another. A graph is called un-weighted if each edge is assigned with equal weight; or weighted if edges in the graph are assigned with different weights. A path in a graph is a sequence of nodes in which each node is connected to the next node in the sequence by an undirected/directed edge. The length of a path is defined as the total number of edges that are traversed along the path in an un-weighted graph, and the sum of the weights of the edges in a weighted-graph, respectively. A sub-graph of a given graph consists of a subset of nodes and the edges that connect between these nodes in the graph (Figure 1.6). Graph theory has important applications in a wide range of scientific problems. For example cellular metabolism has been modeled using graph theory where each metabolite is a node and the biochemical reactions that transform one metabolite to another metabolite are the edges connecting the corresponding nodes. Graph theory has been also applied to studying the Bose-Einstein-Condensation, sociology and the dynamics of Word Wide Web(Albert and Barabasi, 2002).

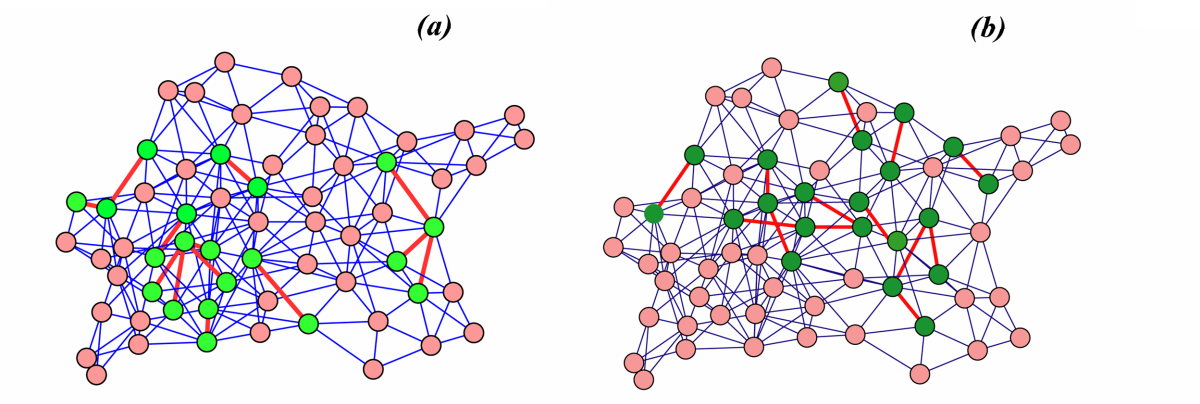

FIGURE 1.6. A sample graph and two sub-graphs are shown where the nodes and the edges of the sub-graphs are highlighted with different colors from the rest of the graphs.

CHAPTER 2

COMBINING HYDROGEN EXCHANGE DATA AND COARSE-GRAINED SIMULATION TO CHARACTERIZE FOLDING INTERMEDIATE OF FAT DOMAIN

Introduction

Focal adhesion kinase (FAK) is a non-receptor tyrosine kinase that is expressed in most tissues and is regulated by integrin-dependent cell adhesion(Parsons, 2003). FAK functions in the control of several important biological processes, e.g. cell migration and apoptosis(Schaller, 2001) and FAK is over-expressed in many forms of cancer. The correct localization of FAK to focal adhesions is required for integrin-dependent regulation of FAK and for FAK to direct tyrosine phosphorylation of downstream substrates. The C-terminus of FAK contains a focal adhesion targeting (FAT) domain, which is a four-helix bundle structural motif that is responsible for subcellular localization of FAK.

The avian FAT domain contains a tyrosine in helix 1 at position 926 (925 in the human and mouse sequences) that is a substrate for phosphorylation by Src. Phosphorylation of Y926 creates a docking site for the SH2 domain of Grb2 (Schlaepfer and Hunter, 1997), and the FAT/Grb2 interaction is one of the mechanisms linking FAK to the Ras/MAPK signaling pathway(Schlaepfer and Hunter, 1997). Structural studies indicate that Src-like tyrosine kinases bind to substrates in a β-strand conformation (Brown et al., 1999; Hubbard, 1997). Further structural studies demonstrated that tyrosine phosphorylated ligands form a β-turn to facilitate binding interactions with the Grb2 SH2 domain (Kuriyan and Cowburn, 1997).

Results from these studies suggest that the region around Y926 must adopt an extended conformation for phosphorylation and association with Grb2. There has been speculation (Arold et al., 2002; Liu et al., 2002; Prutzman et al., 2004) that for phosphorylation of Y926 and subsequent Grb2 binding to occur, the FAT domain must pass through an 'open' conformation in which helix 1 extends from the four helix bundle allowing the region flanking tyrosine 926 to adopt the necessary conformation.

While X-ray crystal (Arold et al., 2002; Hayashi et al., 2002) and NMR structures (Liu et al., 2002; Prutzman et al., 2004) reveal a four helix bundle fold for the FAT domain, a second structure has also been described. One of the reported crystal structures (Arold et al., 2002) is a dimer that has undergone 'domain exchange', in which the N-terminus and helix 1 of one molecule associates with helices 2 through 4 of the other symmetry-related molecule. The formation of the domain-exchanged dimer was speculated to proceed through an intermediate state where helix 1 transiently separates from the core bundle(Arold et al., 2002). The intermediate state of the FAT domain that promotes formation of the domain-swapped dimer may also facilitate phosphorylation at tyrosine 926 and Grb2 binding(Arold et al., 2002). NMR studies of the FAT domain in solution revealed that residues at the carboxy-terminus of helix 1, loop 1, and the amino terminus of helix 2 (residues 941-951) showed line-broadening consistent with conformational exchange on the NMR time scale(Prutzman et al., 2004). The loop region between helices 1 and 2 was consequently identified as a putative 'hinge region' that may be responsible for promoting a partially unfolded intermediate state, similar to the intermediate state that would produce the domain-swapped dimer. However, detection and characterization of such protein intermediate states is often non-trivial.

Although weakly-populated protein-folding intermediates are often difficult to structurally characterize, hydrogen exchange methods have proven to be a powerful technique for identifying and characterizing kinetic and equilibrium folding intermediates (Bai et al., 1995; Chamberlain et al., 1996), but are limited in their ability to describe the structure of the intermediates. By combining hydrogen exchange data with discrete molecular dynamics (DMD)(Dokholyan et al., 1998) simulations, we have been able to capture structural details of the intermediate state ensemble for FAT domain folding, which has allowed us to reconstruct the conformers that we believe are important for FAK signaling through the FAT domain.

Material and methods

Hydrogen exchange protection factor

The relationships between the hydrogen exchange rates, transient unfolding mechanisms, and the structural stability of proteins have been previously described (Bai et al., 1994; Maity et al., 2003). The exchange rates of backbone amide protons are commonly expressed as protection factors, $P_f = k_{rc}/k_{ex}$, where k_{ex} is the experimentally measured hydrogen exchange rate and k_{rc} is the intrinsic rate of hydrogen exchange in the unstructured protein. Protection factors can be calculated using the method presented by Bai and coworkers (Bai et al., 1993). In the EX2 limit, usually satisfied when the structure of the protein is stable, the protection factor is equivalent to the equilibrium constant for the unfolding transition that makes the amide hydrogen exchange-competent. In this case, the protection factors may be used to calculate the free energy of the structural opening event:

$$\Delta G_{HX} = -RT \ln P_f . \qquad (2.1)$$

Here, NMR HX data were collected by Dr. Sharon Campbell's lab(Dixon et al., 2004). The derived protection factors were used as experimental constraints in discrete molecular dynamics simulations of the FAT domain folding process.

Incorporating protection factors into the Gō Model

In brief, the protection factors obtained from hydrogen exchange experiments can provide the free energy associated with the stability of the protein at each amino acid residue, provided that the amide protons are in the EX2 limit (Bai et al., 1994). In the Gō model, an attractive potential is assigned to each native contact, the sum of which gives the total energy of the protein. The free energy difference between the folded and unfolded states of the protein can be determined at each amino acid from the sum of the potentials of the native contacts in which the amino acid residue participates(Dixon et al., 2004). A cost function was used to describe the difference in the free energy values derived from the protection factors and the free energy values calculated by taking the summation over the potentials of the native contacts(Dixon et al., 2004). By using Monte Carlo Metroplis algorithm(Binder, 1995), a minimization of the cost function was then performed to obtain the set of potentials that is most consistent with the experimental protection factors(Dixon et al., 2004).

Discrete Molecular Dynamics Simulations

Interaction model

A discrete molecular dynamics (DMD) algorithm (Dokholyan et al., 1998; Smith et al., 1997; Zhou et al., 1997) was used to study the folding thermodynamics of the FAT domain, as DMD simulations have proven to be very useful in the study of folding kinetics (Borreguero et al., 2002; Ding et al., 2002a; Zhou and Karplus, 1999) and aggregation of proteins (Ding et al., 2002b; Smith and Hall, 2001). The FAT domain was modeled using the 'beads-on-string' method developed by Ding et al. (Ding et al., 2003), with 'beads' corresponding to the C_α, C_β, N, and $C^{'}$ atoms. During the simulation, distance and angle constraints are maintained between the 'bead' atoms. The lowest eergy solution structure, determined by NMR (Prutzman et al., 2004), was used as the native structure and the Gō potential was employed to model the interaction energy between native contacts. The non-bonded interactions V_{ij} were only assigned between C_β atoms (C_α for Gly) of residues *i* and *j* (|*i*-*j*|>2):

$$V_{ij} = \begin{cases} +\infty, | r_i - r_j | \le a \\ \gamma\varepsilon_{ij}, a < | r_i - r_j | < b \\ 0, | r_i - r_j | > b \end{cases}, \qquad (2.2)$$

where $|r_i - r_j|$ is the distance between C_β atoms (C_α for Gly) of residues *i* and *j*. The parameters *a* and *b* are the hard-core diameter (3.25 Å) and the cut-off distance (7.5 Å), respectively. If the C_β atoms (C_α for Gly) for a pair of residues are closer than 7.5 Å in the native state, an attractive potential (γ=-1) was assigned to the pair wise interaction; a repulsive potential (γ=1) was assigned for the interaction between all pairs of residues whose C_β atoms (C_α for Gly)

were separated by more than 7.5 Å in the native state. The depths of the attractive square-well, ε_{ij} in Equation 2.2, are equal in the *unscaled* Gō model, whereas in the *scaled* Gō model, ε_{ij} is assigned different strengths according to experimental data, as described below.

Scaling Gō model potentials using experimental protection factors

Since the measurements of the protection factors were performed under conditions where the folded state is dominant, we expect that hydrogen exchange for most of the residues is governed by the local fluctuation around the native structure. In the EX2 limit, where the protection factors can be used to determine the free energy associated with local protein stability, the measured protection factors can be related to the Gō model interaction energy according to Equation 2.3:

$$\ln P_i \approx \sum_j f_{ij}^F \varepsilon_{ij} \approx \sum_j \varepsilon_{ij} \ , \qquad (2.3)$$

where the summation is taken over all residues *j* forming native contacts with the residue *i,* and f_{ij}^F represents the probability of contact formation between residues *i* and *j* in the native state ensemble. We expect that the values of f_{ij}^F are approximately 1. Thus, in Equation 2.3, the protection factor reflects the heterogeneity in the contact energies.

The experimental protection factors were quantitatively determined for the subset of residues that could be monitored in real time. While some of the remaining residues could be approximated based on the CLEANEX-PM experiments(Hwang et al., 1997; Hwang et al., 1998), we did not calculate protection factors for residues that were not assigned in the ^{1}H-^{15}N HSQC NMR spectrum or for which we could not determine the observed hydrogen

exchange rate using the experiments performed. So, to determine the set $\{\varepsilon_{ij}\}$ that is most consistent with the experimental data, a cost function was constructed that was minimized using Monte Carlo Metroplis algorithm(Binder, 1995):

$$W = \left\langle \left(\ln P_{F_i} - \sum_j \varepsilon_{ij} \right)^2 \right\rangle_{measured} + \left\langle \left(\varepsilon_{ij} - \langle \varepsilon_{ij} \rangle_{all} \right)^2 \right\rangle_{all}, \quad \varepsilon_{ij} \geq 0.1 \qquad (2.4)$$

In the first term, the summation of *j* is taken over residues forming native contacts with *i*. $<...>_{measured}$ is the average taken over all the residues *i* with measured protection factors. For residues that were observed using the CLEANEX-PM experiment(Hwang et al., 1997; Hwang et al., 1998), a small protection factor was assigned with a value of 10, as the results are not sensitive to the exact magnitude of this value. In the second term, $<\ldots>_{all}$ is the average taken over all the native contacts. The purpose of including the second term was to determine the pair-wise contact energies for residues that were not directly constrained by available experimental data, such as contacts between residues in which neither member of the pair has a measured or approximated protection factor. All native contact energies were constrained to have a minimum attraction interaction of 0.1.

Results and discussions

Scaled Gō model

The native contacts that were strengthened in the scaled Gō model are shown in the lower right triangle of Figure 2.1. After generating a set of scaled contact potentials $\{\varepsilon_{ij}\}$ (as described in Methods), we verified that our set was consistent with the experimental protection factors. Therefore, the contact frequencies, f_{ij}^F, from the trajectories at low

simulation temperature, where the folded state is highly populated, were combined with the potentials $\{\varepsilon_{ij}\}$ set according to Equation 2.3, to produce calculated protection factors.

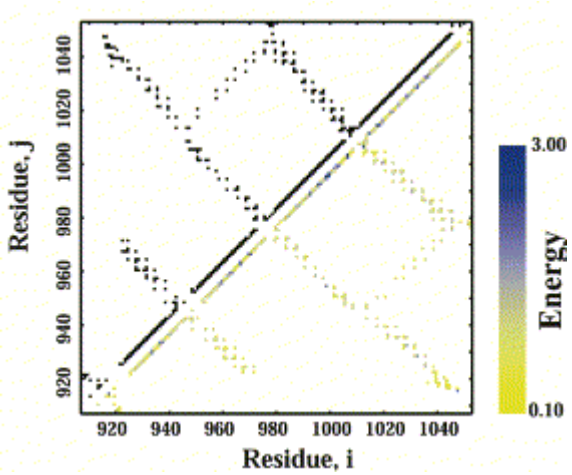

FIGURE 2.1: Energy map for the FAT domain. Contact map for the native state of the FAT domain with the pair wise contact potentials, scaled from the HX protection factors, shown in the lower right triangle. The energy scale for the contact potentials is in units of ε.

The calculated protection factors show a strong correlation to the experimental protection factors for the FAT domain (Figure 2.2) with a correlation coefficient of 0.99. In contrast, if the set of unscaled contact potentials is used instead of the scaled set, the correlation coefficient is only 0.07. The set of scaled contact potentials produced from the Monte Carlo minimization of the cost function (Equation 2.4, see Methods) therefore represents an accurate incorporation of the experimental protection factors into the DMD simulations.

To test the robustness of our method for scaling contact potentials, calculated protection factors were generated for three other proteins whose hydrogen exchange protection factors are available in the literature; barnase (Perrett et al., 1995), horse heart cytochrome C (Milne et al., 1998), and ribonuclease H (Chamberlain et al., 1996). Since the protection factors were used solely as adjustable parameters for our scaling model, it was not important how the

exchange experiments were conducted or whether the EX2 limit was satisfied. Contact energies for each of these proteins were scaled according to their experimental protection factors, and then reconstituted in the form of calculated protection factors. Strong correlations were observed between the experimental and calculated protection factors (Figure 2.2) for all three proteins, with correlation coefficients similar to that obtained for the FAT domain. These results verify that our method for scaling the contact potentials is self-consistent.

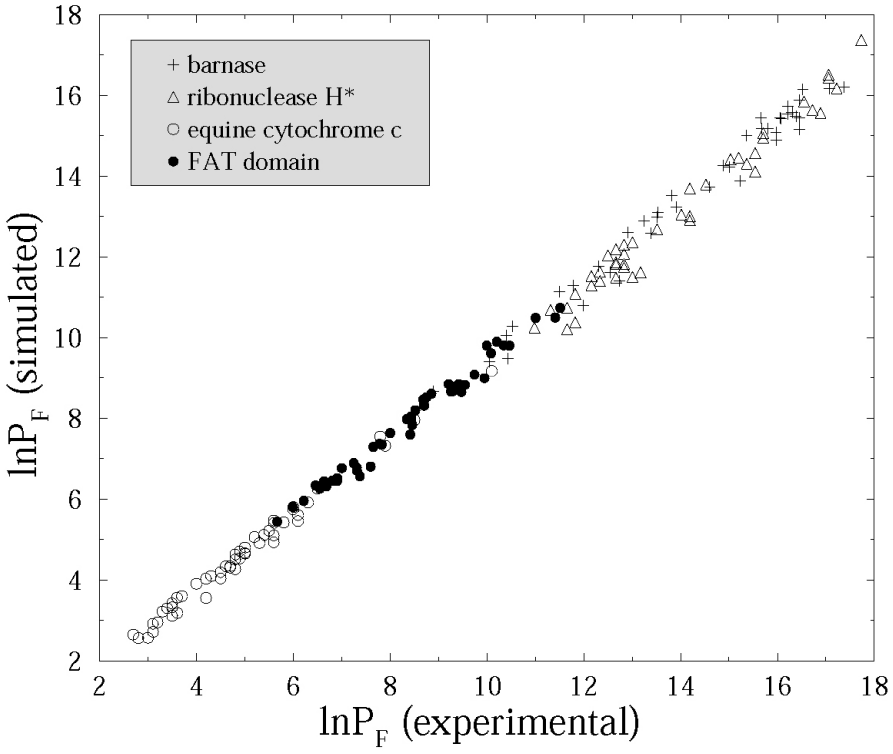

FIGURE 2.2: Correlation between experimental and calculated protection factors. Correlation plot comparing the experimentally determined protection factors with the reconstituted protection factors after the Monte Carlo scaling of the pair wise contact potentials of the FAT domain. The algorithm was applied to three other well characterized proteins: barnase, mutant ribonuclease H, and equine cytochrome c for comparison.

Recently, Vendruscolo and coworkers (Vendruscolo et al., 2003) have proposed a phenomenological approach to utilize hydrogen exchange data to bias the sampling in conformational space toward rare fluctuations of native proteins. In contrast, our method is intended to refine the interaction model with HX data so that we can better characterize the energetics underlying the folding process. Using this method, we are able to reconstruct particular ensembles at a given temperature and also characterize the thermodynamics and kinetic parameters of the FAT domain based on the rapid DMD simulation algorithm.

The temperature dependence of the average potential energy, derived from the scaled Gō model simulation of the FAT domain, is shown in Figure 2.3. At low temperatures ($T << 1$, where the temperature is in reduced units of ε/k_B), the FAT domain exists predominately in its native folded state, whereas at high temperatures ($T > 1$), it is present mostly in an unfolded state. The sigmoidal curve shows a large increase of potential energy with the increase of temperature, which indicates a highly cooperative step in the transition. The shape of the curve alone does not provide the number of states there are at a given temperature. Rather, it reflects the average potential energy over all states. To determine which states are present, we analyzed the distribution of potential energy states.

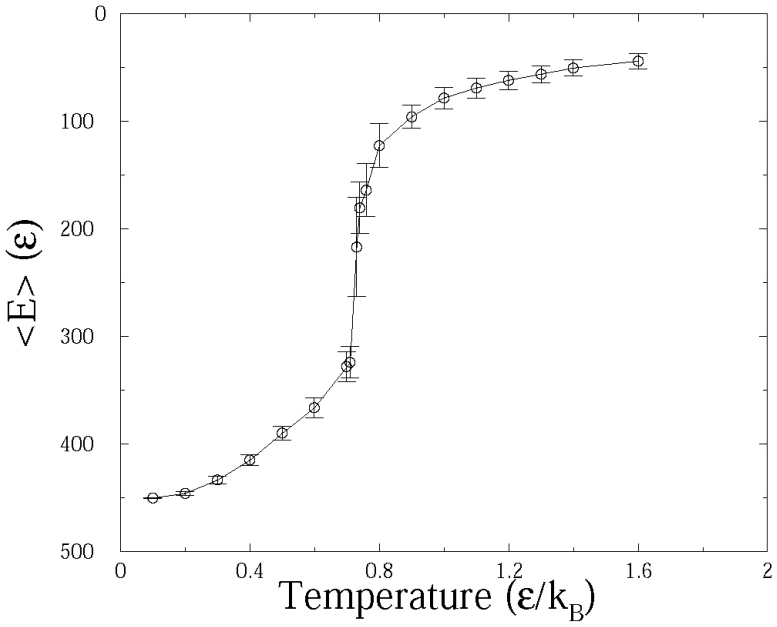

FIGURE 2.3: The folding trajectory for the scaled DMD simulations. The folding trajectory of the FAT domain from DMD simulations using the Go model with the pairwise contacts scaled to agree with the experimentally determined protection factors. Error bars are shown depicting the standard deviation. The temperature of the simulation and the total energy of the protein are shown in reduced units based on the potential energy of the protein's native contacts, ε.

The probability distribution associated with the total potential energy of the protein is shown at three different simulation temperatures in Figure 2.4. At T=0.68 (bottom panel) the distribution of states is concentrated in a single distribution, with an energy value near the native folded state of the protein. As the simulated temperature is raised to 0.72 (middle panel), three distinct distributions are apparent: a native-like folded (F) state, an intermediate (I) state, and a largely unfolded (U) state. Near the end of the transition (top panel), T=0.81, a single distribution is observed with an energy that is near the completely unfolded state of the protein. The presence of an intermediate state is clearly detected in these distributions

and a Gaussian fitting, used to determine the relative distribution of the states, shows that the intermediate state is significantly populated near the midpoint of the transition.

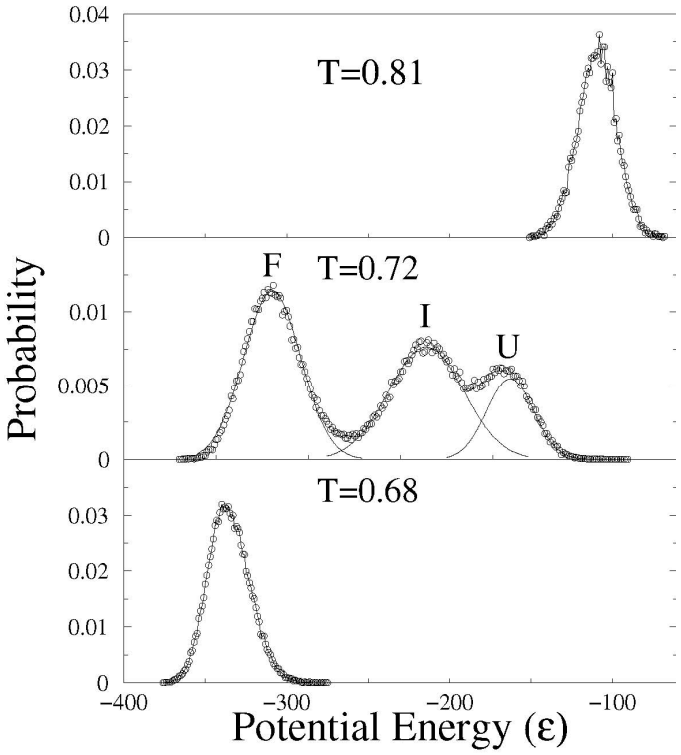

FIGURE 2.4: The probability distribution at three simulation temperatures. The probability distribution of the total potential energy for the FAT domain, based on the summation over all pair wise contacts, for three simulated temperatures along the folding transition. At T = 0.68 (bottom panel) the distribution of states is concentrated in a single distribution, with an energy value near the native, folded state of the protein. As the simulated temperature is raised to 0.72 (~T_M, middle panel), three distinct distributions are apparent: a native-like folded (F) state, an intermediate (I) state, and a largely unfolded (U) state. A Gaussian fitting was used to determine the relative distribution of the three states. Near the end of the transition (top panel), T = 0.81, a single distribution is observed with an energy that is near the fully denatured state of the protein.

The FAT folding intermediate

The intermediate state is not a discrete structure, but an ensemble of conformations that have potential energies that are distinct from the ensembles of the folded and unfolded states. There are two main characteristics of the intermediate detected by the scaled Gō model simulations of the FAT domain: 1) helix 1 separates from the helix bundle and 2) helix 1 loses helical structure. The frequency map for the intermediate state at T=0.72 is shown in the lower right triangle of Figure 2.5. The frequency map indicates the probability that a contact that exists in the native state, is being made at the given temperature of the simulation. Inspection of the intra-helix (along the diagonal) and inter-helix (perpendicular to the diagonal) contacts, indicates that contacts made by helix 1 (residues 924-943) occur with lower frequency than the contacts made by residues in the other helices.

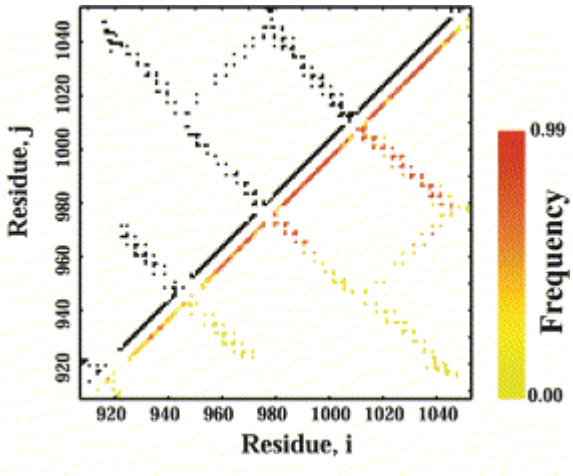

FIGURE 2.5: Frequency map for the intermediate ensemble. Contact map for the FAT domain with the lower right triangle showing the frequency of the contacts in the intermediate ensemble at T=0.72.

A single FAT domain molecule from the domain-swapped dimer structure (Arold et al., 2002) is shown next to a snap shot of the folding intermediate ensemble produced by the scaled Gō simulations in Figure 2.6; both structures share the characteristic of having helix 1 extended from the helix bundle. The intermediate structure is weakly populated at temperatures that stabilize the four helix bundle fold, but increases in the DMD simulations when the temperature of the simulation is raised sufficiently. The domain swapped dimer structure was produced from crystals at room temperature (Arold et al., 2002), which suggests that helix 1 transiently extends from the helix bundle at temperatures well below the melting temperature (T_M = 83.4 at pH = 6.0, unpublished) and, consequently, allows the domain exchanged dimer to form. Crystals of the domain-swapped dimer only appeared after three months, leading the authors to speculate that helix-exchanged molecules represent a minor population of the FAT domain. Although a domain-swapped dimer form of the FAT domain has been characterized, there is little evidence to support a role for a FAK dimer *in vivo*(Arold et al., 2002). As such, the FAT domain swapped dimer may be a byproduct of the FAT domain sampling a more 'open' intermediate which becomes populated at concentrations used for X-ray crystallization. In contrast, mounting evidence suggests that helix 1 is conformationally dynamic, and conformational changes in helix 1 are important for phosphorylation of Y926 by Src, recognition by Grb2, initiation of MAPK signaling, subcellular localization of FAK, and focal adhesion turnover.

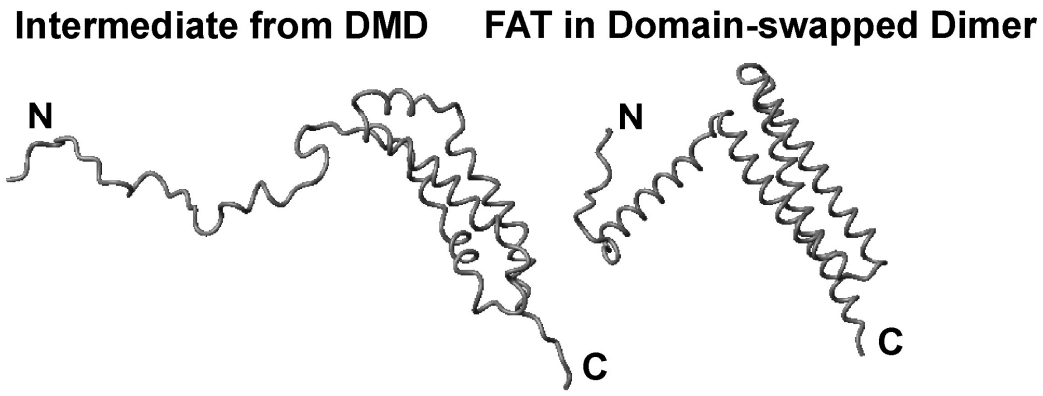

FIGURE 2.6: Comparison of the FAT domain intermediate from DMD simulations with the domain-swapped dimer. DMD simulations on the FAT domain of FAK produce an ensemble of structures for the folding intermediate. A single representative structure from the ensemble is shown for comparison with a single molecule from the domain-swapped dimer crystal structure (1K04) of the FAT domain. In both structures, helix 1 has separated from the helix bundle and in the intermediate structure this is accompanied by the loss of helical character in helix 1. The figure was prepared using MOLMOL.

The scaled Gō model simulations of the FAT domain reveal a folding-intermediate ensemble in which helix 1 is less structured and separates from the helix-bundle. Although the HX data indicates that residues in helix 1 are less protected than residues in helices 2-4, the HX data alone cannot confirm the presence intermediate state. Hydrogen exchange experiments as a function of chemical denaturant are able to detect cooperative units of unfolding (Bai et al., 1994) and subglobal unfolding, and have been employed to demonstrate sub-global unfolding in the four-helix bundle protein, cytochrome b_{562}(Fuentes and Wand, 1998). We recognize that our simulations produce a subglobal opening in the FAT domain which denaturation studies may be able to experimentally confirm. However, identifying the contributions to the effective hydrogen exchange rate (i.e.: local, subglobal, and global) was

not necessary for our method, since scaling the native contacts was based solely on the effective exchange rate. Furthermore, using our scaled Gō model simulations, we were able to both identify cooperative units of unfolding and reconstruct coarse structures of the intermediate state.

The general features of this intermediate state are consistent with our recent solution structural studies of the FAT domain, in which line-broadening associated with NH resonances in and around helix 1 and 2, led us to propose that a proline-rich 'hinge region' in the FAT domain produces strain resulting in enhanced conformational dynamics of helix 1 (Prutzman et al., 2004). Moreover, the intermediate state ensemble detected by our DMD/HX approach is consistent with the FAT domain 'sampling' a more open state that leads to formation of a domain-swapped dimer(Arold et al., 2002). The region of the FAT domain that is exposed in this intermediate state ensemble, i.e., helix 1, has been shown to play a role in FAK phosphorylation, paxillin binding and FAK localization.

The Role of the Open and Closed Conformers of the FAT Domain

Results from this study combined with our previous observations (Prutzman et al., 2004) support the existence of 'open' and 'closed' conformations of the FAT domain. The four-helix bundle (closed conformation) of the FAT domain appears important for two distinct, yet related processes: paxillin binding and targeting FAK to focal adhesions. Recent structural and biochemical studies have elucidated paxillin binding interactions with the FAT domain (Gao et al., 2004; Hayashi et al., 2002; Hildebrand et al., 1995; Hoellerer et al., 2003; Tachibana et al., 1995). However, the process by which FAK becomes localized to focal adhesions is not completely understood. Although the binding of paxillin and focal adhesion

targeting of FAK show a strong correlation (Chen et al., 1995; Tachibana et al., 1995), studies of FAK mutants have revealed that paxillin-binding is dispensable for localizing FAK to focal adhesions (Cooley et al., 2000). It has therefore been suggested that paxillin-binding represents one mechanism for localizing FAK to focal adhesions and that alternative mechanism(s) exist (Cooley et al., 2000). One possible alternative mechanism involves binding of the focal adhesion protein talin (Chen et al., 1995) to the FAT domain, which may facilitate the targeting of FAK to focal adhesions. Interestingly, mutational analyses indicate that while paxillin requires an intact FAT-structure for binding, talin does not (Chen et al., 1995; Hayashi et al., 2002). Therefore, the 'closed form' of the FAT domain may be important for targeting FAK to focal adhesions through a paxillin-binding mechanism.

Recent structures of the FAT domain complexed to paxillin-derived LD peptides indicate that paxillin binds the FAT domain at interfaces between helices 1 and 4 as well as between helices 2 and 3 (Gao et al., 2004; Hoellerer et al., 2003; Liu et al., 2002). Moreover, paxillin requires a rigid 4-helix bundle conformation for FAT binding but does not induce significant structural rearrangements in the FAT domain upon binding (Gao et al., 2004; Hoellerer et al., 2003; Liu et al., 2002). The two sites of FAT involved in paxillin LD peptide binding have been designated hydrophobic patch 1 (HP1) and 2 (HP2) (Gao et al., 2004; Hoellerer et al., 2003). One of the binding sites, HP2, partially obstructs Y926 (Hayashi et al., 2002; Hoellerer et al., 2003). Not only is paxillin-binding likely to interfere with phosphorylation of Y926, but mounting evidence suggests that phosphorylation and Grb2 binding to the FAT domain requires structural rearrangements within helix 1 of the FAT domain (Arold et al., 2002; Liu et al., 2002). However, structural rearrangements within helix 1 are likely to perturb the paxillin binding site at HP2, which lies between helices 1 and 4, so the

conformational changes required for phosphorylation at Y926 are likely to disrupt paxillin FAT/HP2 interactions. Therefore, paxillin binding at HP2 and phosphorylation of Y926 are likely to be mutually exclusive events.

A previously reported double mutant, V955A/L962A, showed disruption of paxillin binding but retained the ability to target focal adhesions (Cooley et al., 2000). While the V955A/L962A double mutant does not cause significant structural alterations in the FAT domain (Prutzman et al., 2004), the mutations in helix 2 have been predicted to cause subtle perturbations in the 4-helix bundle core of FAT by disruption of specific hydrophobic interactions between helices 2 and 3. Our labs have shown that this double mutant exhibits an approximately 8-fold increase in dimerization (by gel-filtration) and a dramatic increase in Y926 phosphorylation *in vivo* (Prutzman et al., 2004). The increase in dimerization suggests that the protein is more frequently sampling 'open' conformations that expose hydrophobic residues within the amphipathic helices (Prutzman et al., 2004). In the same study, it was shown that helix 1, when expressed as a GST fusion protein, was phosphorylated approximately 8-fold more than a GST-FAT domain fusion protein (Prutzman et al., 2004). We have previously postulated that the increased phosphorylation, dimerization, and disruption of paxillin-binding observed for V955A/L962A FAT variant may be the result of conformational dynamics involving helix 1 (Prutzman et al., 2004).

We have proposed a model for the role of FAT conformational dynamics in FAK-mediated cell adhesion and signaling processes, based on the available structural, biochemical and biological information on FAK combined with structures produced from the scaled Gō model simulations. In our model, the open conformation of the FAT domain is important for FAK signaling. Phosphorylation of Y926 creates a consensus recognition site

(pYNQV) for the SH2 domain of Grb2 (Rahuel et al., 1996), However, phosphorylation and subsequent binding requires a more extended conformation of helix 1 (Rahuel et al., 1996), and therefore helix 1 must have some degree of conformational flexibility to accommodate the required structural rearrangements (Arold et al., 2002; Liu et al., 2002). Grb2 binding at pY926 initiates signaling through the Ras/MAPK pathway. Therefore, transient population of the intermediate state conformers could favor the structural rearrangements of helix 1 that facilitate Grb2 activation of the MAPK pathway.

We also postulate that the open conformation aids in the removal of FAK from focal adhesions, as phosphorylation of Y926 has been linked to exclusion of FAK from focal adhesions (Katz et al., 2003). Moreover based on current structural data, it has been speculated that the SH2 domain of Grb2 can not bind phosphorylated Y926 while FAK is still associated with the focal adhesion complex (Liu et al., 2002). Although the precise sequence of events is not known, we speculate that the open conformer of FAT, once phosphorylated, initiates FAK-mediated MAPK signaling and also promotes the removal of FAK from focal adhesions.

CHAPTER 3

CHARACTERIZING THE LARGE-SCALE STRUCTURAL DYNAMICS OF VINCULIN RELATED TO ITS ACTIVATION

Introduction

Vinculin, a highly conserved 117 kDa intracellular protein (1066 residues), plays critical roles in the maintenance and regulation of cell shape, adhesion and migration that are essential to many physiological and pathological processes such as embryogenesis, wound healing and metastasis(DeMali et al., 2002; Fernandez et al., 1992; Fernandez et al., 1993; Volberg et al., 1995; Xu et al., 1998). In its inactive state, vinculin is held in a "closed", auto-inhibited conformation by intra-molecular interactions between the head and tail domains (Vt)(Johnson and Craig, 1995a; Johnson and Craig, 1995b). Upon recruitment to cell-cell and cell-matrix junctions, vinculin becomes activated and provides a structural link between the membrane and actin filaments in both locales(Pokutta and Weis, 2002; Zamir and Geiger, 2001; Critchley, 2000), although the exact role of vinculin in the formation of adhesion complexes is unknown. The crystal structure of vinculin is composed of a series of helix bundles that are arranged into five distinct domains(Bakolitsa et al., 2004) (D1-D5, Figure 3.1). Each of the first three domains (D1-D3) comprises two four-helix bundles that share a central long α-helix, whereas D4 is a single four-helix bundle that is connected to Vt (D5) by a proline-rich linker region (Figure 3.1). Domains D1 and D3 form a pincer-like structure holding the vinculin tail in an auto-inhibited state (Figure 3.1) in which many ligand-binding

sites are occluded. Domain D2 stabilizes the pincer-like structure by forming extensive contacts with domain D3.

The activation of vinculin requires the release of the interaction between the D1 (residue 1-258) and Vt (residue 896-1066) domains, which is triggered by binding to different ligands. Early biochemical and structural studies established the role of an acidic phospholipid PtdIns(4,5)P_2 in disrupting the interaction between D1 and Vt by binding to Vt domain(Gilmore and Burridge, 1996; Weekes et al., 1996; Bakolitsa et al., 1999). The release of the D1-Vt interaction further exposes cryptic binding sites in vinculin for other ligands, including talin, α-actinin, α-catenin, vasodilator-stimulated phosphoprotein (VASP), vinexin, ponsin, actin-related protein complex (Arp2/3), paxillin and actin(Zamir and Geiger, 2001). In contrast, the role of talin in releasing the head-tail interaction was uncovered more recently(Izard and Vonrhein, 2004; Izard et al., 2004). It was found that the binding of specific short talin peptides (~30 amino acids) to the D1 domain alone is sufficient for releasing intra-molecular head-tail interactions in the D1-Vt complex by provoking significant structural changes of the D1 domain, suggesting an alternative pathway of vinculin activation, in which PtdIns(4,5)P2 may not be required(Izard and Vonrhein, 2004; Izard et al., 2004). However, Bakolitsa et al.(Bakolitsa et al., 2004) and Gilmore and Burridge(Gilmore and Burridge, 1996) demonstrated that larger fragments of talin bind poorly to full-length vinculin *in vitro*. When PtdIns(4,5)P2 is added, binding strength increases four-fold(Gilmore and Burridge, 1996). Hence it is more likely that the activation of vinculin is achieved by a combinatorial binding of ligands rather than any single ligand(Bakolitsa et al., 2004; DeMali, 2004)

Despite intensive biophysical and biochemical studies, a global and dynamic picture of vinculin activation is still lacking. Experimental characterization of the large-scale conformational dynamics relevant to vinculin activation is challenging. Even using computational methods, the time-scale of the conformational dynamics and the size of vinculin are beyond the resolution capabilities of the current all-atom molecular dynamics simulation techniques. Hence, to uncover the large-scale conformational dynamics associated with vinculin function, we utilize rapid discrete molecular dynamics (DMD)(Dokholyan et al., 1998; Smith et al., 1997; Rapaport, 2004) techniques. We find distinct and complementary roles of internal (inherent flexibility, domain-domain interactions within vinculin) and external (talin binding) factors in allosteric control of vinculin, suggesting possible mechanisms for vinculin activation.

Materials and methods

Homology modeling of missing loop structure

The conformation of a loop (856-874) is missing in the X-ray structure of vinculin. We use the biopolymer module of the program SYBYL to reconstruct the backbone conformation of this loop. The reconstruction is based on homology modeling method(Sanchez and Sali, 1997). By searching the structural homologues of this loop in the Brookhaven Protein Databank (PDB)(Berman et al., 2000), we use SYBYL to generate a candidate list of loop structures for reconstruction. From this list as the template for reconstruction we choose the loop structure (PDB accession code: 1KZQ, chain B, residues 191-212) that has the highest sequence identity (26.1%) to the missing loop. We further use SYBYL to integrate the reconstructed loop conformation into the original vinculin structure.

We further add side chains to amino acids on the reconstructed backbone. We determine the optimal rotamer states of side-chains by a Monte-Carlo minimization procedure(Ding et al., 2006).

Protein and interaction model

We perform DMD simulations using a simplified two-bead protein model, in which each residue is represented by one backbone bead C_α and one side-chain bead C_β (only C_α for Gly). The detailed implementation of covalent bonds and constraints that maintain the geometry of each residue in the model can be found in Ref. (Ding et al., 2002a). In addition to the covalent bonds and constraints, we use the Gō potential(Abe and Go, 1981; Go and Abe, 1981) to model the non-bonded interactions within monomers. The chicken vinculin crystal structure(Bakolitsa et al., 2004) (PDB accession code 1ST6) with reconstructed loop is used as native structure to assign the Gō potential. We also incorporate backbone hydrogen bonding interaction into the simulations (Ding et al., 2002b).

Thermodynamic and kinetic simulations

In thermodynamic studies, prior to the equilibrium simulations we perform simulations for $1x10^5$ time units from the initial temperature T=0.1 to various target temperatures in the range between T=0.1 and T=2.0 (simulation temperature is in units of ε/k_B, where ε is the energy unit and k_B is Boltzmann's constant). We then perform equilibrium simulations for $1x10^6$ time units at corresponding target temperatures. In kinetic studies, we perform twenty unfolding simulations starting from the same native structure of vinculin but different initial velocities. In each kinetic unfolding simulation, we gradually (during $5x10^5$ time unit)

increase the system temperature from T=0.1 to T=0.9. We then analyze the pattern of dissociations between domains and within domains during the course of unfolding.

Fraction of native contacts as a measure of dissociation magnitude

A native contact is defined to exist in a given conformation if the C_β atoms of two residues are within a cutoff distance (7.5 Å) both in this conformation and in the native structure. The cutoff distance used to define the native contact is the same as the one used to define structure-based Gō potential. The fraction of native contacts (Q) is defined as the ratio between the number native contacts in a given conformation and in the native structure. It takes the value ranging from one (when a protein adopts native structure) to zero (when a protein is fully-unfolded) and has been used as a reaction coordinate in the study of protein folding(Onuchic et al., 2000).

Characterization of the principal motions near the native state

We use essential dynamics to characterize the principal motions near the native state. The essential dynamics(Amadei et al., 1993; Ichiye and Karplus, 1991) is based on the diagonalization of the covariance matrix constructed from fluctuations of C_α atoms in the simulation trajectories in which the overall translation and rotation have been removed:

$$M = < (X_i - X_{i,0})(X_j - X_{j,0}) > \qquad (3.1)$$

X_i (X_j) in Eq.(3.1) are the separate x, y, z coordinates of the ith (jth) C_α atom (i,j=1...N, N is the total number of C_α atoms) fluctuating around its average $X_{i\,0}$. The average is taken over all the snapshots along the trajectories used for calculation. We use the trajectory from DMD simulation at low temperature (T=0.2) where the native state is the most favorable state and

the protein undergoes conformational fluctuations near native state. The diagonalization of Eq.(3.1) yields a set of eigenvectors (describing directions in the high-dimensional configurational space) and eigenvalues (represent the mean square fluctuation of the total displacement along the eigenvectors). The first few eigenvectors with the highest eigenvalues describe principal motions. Motions along these eigenvectors are mainly large anharmonic fluctuations and generally can be linked to the biological functions of the proteins. The motions described by eigenvectors with small corresponding eigenvalues represent harmonic (Gaussian) fluctuations, which are thermal fluctuations in nature.

Simulation of the interaction between vinculin binding site peptide from talin and vinculin

To investigate the role of talin binding in vinculin activation, we perform simulations of the binding of vinculin binding site (VBS1, residue 605-636) from talin to vinculin. We model the interaction between vinculin and VBS1 by constructing the following effective potential:

$$\begin{aligned} E_{tot} &= E_{vinculin-free} + E_{VBS1-free} + E_{interaction\ (vinculin-VBS1)} \\ &= \sum_{ij} \Delta_{ij}^{vinculin-free} + \sum_{i,j} \Delta_{ij}^{VBS1-free} + (\alpha \sum_{i,j} \Delta_{ij}^{D1-bound} + \sum_{i,j} \Delta_{ij}^{D1-VBS1}), \end{aligned} \quad (3.2)$$

where $\{\Delta_{ij}^{vinculin-free}\}$ and $\{\Delta_{ij}^{VBS1-free}\}$ are the Gō-like pair-wise contact energies corresponding to the free state of vinculin(Bakolitsa et al., 2004) (PDB accession code 1ST6) and the free state of VBS1(Papagrigoriou et al., 2004) (PDB accession code 1SJ7), respectively. Matrices $\{\Delta_{ij}^{D1-bound}\}$ and $\{\Delta_{ij}^{D1-VBS1}\}$ are the Gō-like contact energies of the VBS1-bound state of D1 domain and between D1 domain and VBS1 that are assigned based on the D1-VBS1 complex structure(Papagrigoriou et al., 2004) (PDB accession code 1TO1). The inclusion of contact energies for the VBS1-bound state of D1 domain serves to effectively account for the conformational change of D1 upon VBS1 binding. For units that dissociate from each other,

it is important to scale effective energy contribution to reflect the additional translational entropy gained by interacting domains. Parameter α is such scaling coefficient, the range of which is determined by consistency of simulations with experimental observations. More specifically, we determine the range of α by satisfying the following two conditions to be in agreement with experimental observations: first, without VBS1, the native state of the free D1 is the most stable state (RMSD from free D1 and VBS1-bound D1 structure are <2.4Å and >3.0Å, respectively); second, with VBS1, the experimentally-determined D1-VBS1 complex is the most stable state at low temperatures (RMSD from complex structure is <2.4Å). With the determined range of α between 0.65 and 0.72, we then perform simulations in the presence of both VBS1 and vinculin at various temperatures. The simulation results are not sensitive to the exact value of α within the determined range. To improve the efficiency of simulation, we constrain the distance between C_β atoms of residue 616ALA of VBS1 peptide and residue 15PRO of D1 domain within 12 Å (these two atoms are within 7.5 in the native D1-Vt complex) so that the VBS1 and vinculin are spatially close to each other.

Results and discussion

The principal motions near the native state

The principal motions that dominate conformational fluctuations in the vicinity of the native state of proteins often recapitulate the structural dynamics underlying their biological functions(Berendsen and Hayward, 2000). To characterize the principal motions near the native state of vinculin, we first perform DMD simulations of vinculin at low temperature, where the native state of the protein is thermodynamically most favorable and the protein mainly undergoes small conformational fluctuations near its native state (Methods). We then

use essential dynamics analysis(Amadei et al., 1993; Ichiye and Karplus, 1991) (Methods) to find the principal motions of the polypeptide chain. The first dominant mode is characterized by a breathing motion between two elements of vinculin structure (Figure 3.1a). One such element consists of the end of the proline-rich linker region (Figure 3.1) and the structural region from Vt helix bundle that is spatially proximal to the proline-rich linker region. The other element contains the structural region from D3 that is in spatial proximity to Vt and one of the two helix bundles from D2 that has extensive contacts with D3. The first dominant mode also involves a global twisting motion that occurs between two helix bundles within D2. The second dominant mode is highlighted by a "holding" and "releasing" motion between Vt domain and pincer-like structure formed by D1, D2 and D3 (Figure 3.1b). The "holding" and "releasing" is synergized between the structural region from the Vt that is spatially proximal to D1, the structural region from D3 that is close to Vt and the middle region in the D2 structure (Figure 3.1b). The third dominant mode involves the nearly-parallel rotation and twisting of D1, D2 and Vt (Figure 3.1c), reflecting the flexibility along the orthogonal directions distinct from the first two dominant modes. The interaction interface between D1 and Vt remains rigid in all three dominant modes of principal motions, indicating that the intra-molecular interactions between D1 and Vt are kept intact under the conformational fluctuations near the native state.

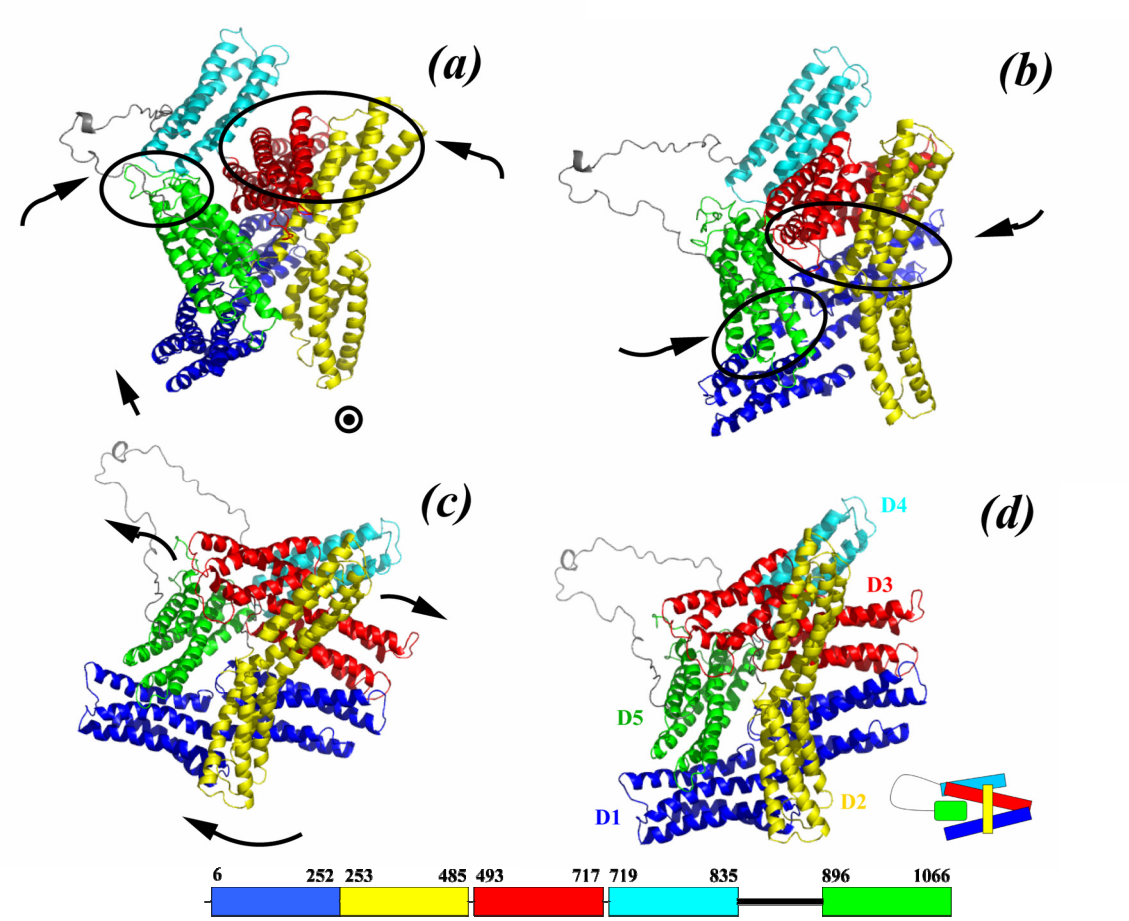

FIGURE 3.1. The native structure of full-length vinculin is shown in the left of Figure 1. It comprises of five distinct domains and one proline-rich linker region connecting the fourth and the fifth domain: D1, 6–252 (blue); D2, 253–485 (yellow); D3, 493–717(red); D4, 719–835 (cyan); proline-rich linker region (838-890) consists of a proline-rich region (838-878) and a "strap" (878-890); D5 (vinculin tail), 896–1066 (green); The pincer-like structure formed between D1, D2 and D3 is illustrated in the bottom-right beside the native structure. All the protein structures are made with PYMOL. The first three dominant modes of principal motions near the native state of vinculin are illustrated as follows: *a*, the first dominant mode; *b*, the second dominant mode; *c*, the third dominant mode.

The thermal "melting" of intra and inter-domain interactions

To further characterize inter- and intra-domain structural plasticities of vinculin that contribute to its function, we study the thermal "melting" of inter and intra-domain interactions in vinculin. We perform equilibrium simulations at various system temperatures ranging from T=0.1 to T=2.0. At each temperature, we use the fraction of native contacts (Methods) to quantify the magnitude of dissociations between and within domains. We plot the fractions of inter-domain and intra-domain native contacts as functions of temperature (Figure 3.2). We find that dissociations between most domains (D1-Vt, D3-D4, Vt-D3, D2-D3, Vt-D4, D1-D3) exhibit cooperative changes with increasing temperatures. In contrast, the association between D1 and D2 domain shows a gradual, rather than cooperative, decrease when temperature increases (Figure 3.2a). In addition, we find that the associations between Vt and D3 domain are completely lost at the temperatures where the majority of the contacts between other domain pairs are kept, indicating that the effective interaction between the Vt-D3 pair is weaker compared to other pairs (Figure 3.2a). Noticeably, the significant dissociations between the tail domain and the "pincer" formed by D1-D3 and between the constituent domains of the "pincer" occur in the temperature range near the midpoint ($T \approx 0.75$) of thermal denaturing curve of the whole vinculin (curve not shown). This observation indicates that the interactions between these domains contribute cooperatively to the stability of the whole vinculin.

As temperature increases, the dissociations between domains are followed by the significant unfolding of individual domains, suggesting that the thermal "melting" of vinculin occurs in a hierarchical manner. We observe a sigmoidal decrease in the fraction of intra-domain native contacts ($Q_{intra\text{-}domain}$) with increasing temperature (Figure 3.2b). For each

domain, at low temperatures most native contacts are kept and, thus, the $Q_{intra\text{-}domain}$ is close to 1 ($Q_{intra\text{-}domain}$ = 1 in the fully folded state). At high temperatures most native contacts are disrupted and the $Q_{intra\text{-}domain}$ is approximately 0 ($Q_{intra\text{-}domain}$=0 in the fully-unfolded state). The midpoint of each curve is the folding transition temperature of the corresponding domain. We find that Vt domain shows higher stability than other domains, which is indicated by the relatively higher folding transition temperature (Figure 3.2b). The remaining domains show different, but comparable, folding transition temperatures.

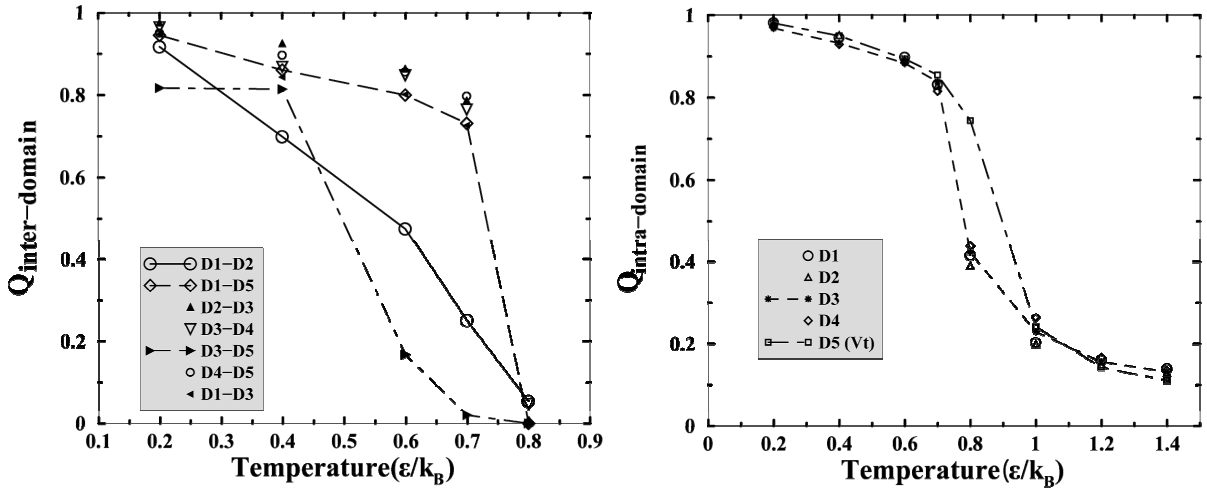

FIGURE 3.2. The thermal melting curves of *a*, inter-domain interactions and *b*, intra-domain interaction are shown. The magnitude of inter and intra-domain dissociations are quantified using the fraction of native contacts made within ($Q_{intra\text{-}domain}$) or between ($Q_{inter\text{-}domain}$) domains. The Q values are calculated using trajectories from equilibrium simulations and plotted as a function of temperature. The higher the Q, the smaller is the magnitude of dissociation.

The kinetics of vinculin unfolding

To better understand how external perturbations lead to the sequential disruption of inter-domain interactions of vinculin, we perform kinetic unfolding of vinculin by gradually increasing the temperatures in the simulation. We collect and analyze twenty kinetic unfolding simulation runs starting from the same native structure of vinculin but different initial velocities (Methods). We find a specific sequential unfolding pattern across different unfolding trajectories (Figure 3.3 and 3.4). Domains Vt and D3 are the first two domains that completely and irreversibly dissociate from each other, which is followed by Vt and D4 dissociation. Next, D1 and D2 dissociate. The native contacts between D1 and D2 are kept at a residual level ($Q_{inter\text{-}domain}$<0.2) due to oscillatory gain and loss of the local interactions between D1 and D2. Further, a series of cooperative dissociations between domains D1-Vt, D2-D3, and D1-D3 (Figure 3.3 and 3.4) occur. The dissociations between these domains occur close to each other in time, although in distinct temporal order in different simulations, suggesting a close interplay between these domains in response to the external perturbation through the unfolding process. Last, a marked decrease of native interactions between D3 and D4 occurs.

Accompanying the inter-domain dissociations, the constituent domains start to unfold. Domains D1, D2 and D4 unfold first, followed by the unfolding of domains D3 and Vt (Figure 3.4). Interestingly, the simulation reveals that the Vt domain unfolds via an intermediate state, whereas the unfolding of other domains exhibit apparent two-state behavior. Further structural characterization of the intermediate state of the Vt domain and the data from NMR experiments suggest its role in Vt function (unpublished data).

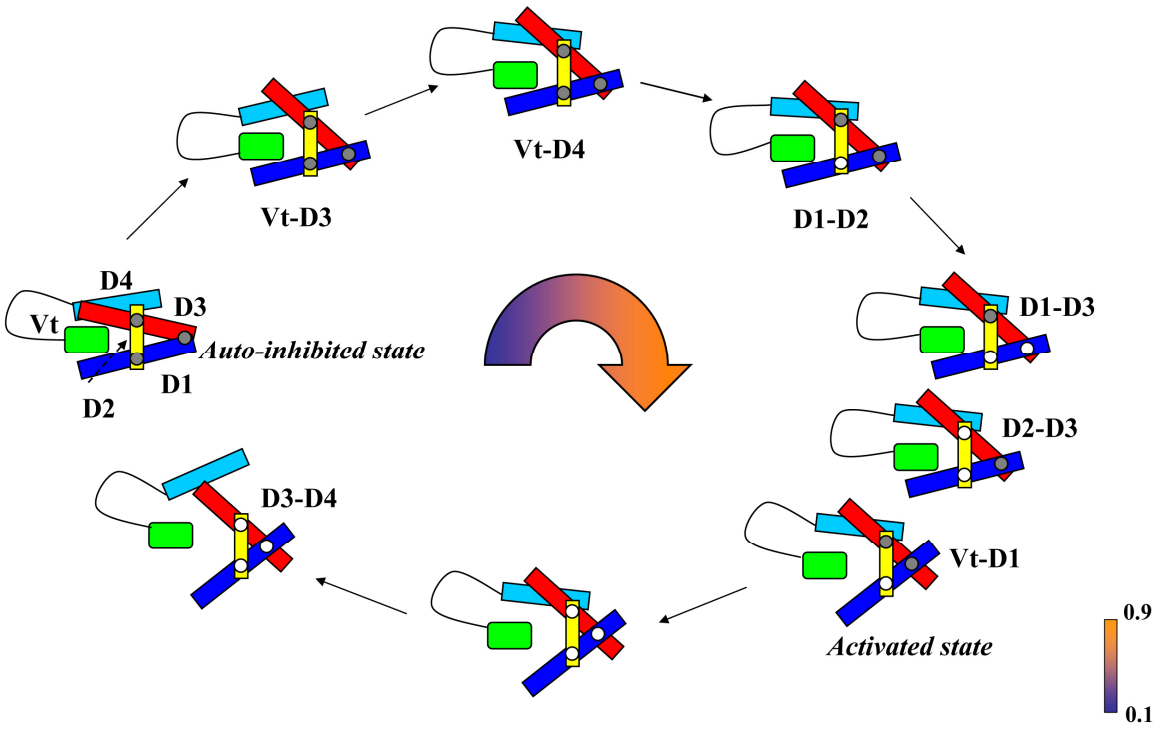

FIGURE 3.3. A cartoon representation is used to illustrate the sequential unfolding events in the kinetic simulations. The five domains are colored as follows: D1, blue; D2, yellow; D3, red; D4, cyan; proline-rich linker region, black; Vt, green. The native contacts between domains in the pincer-like structure formed by D1-D3 are illustrated as screws. The loss of these native contacts is therefore represented as the taking-off of the screws. The unfolding events are connected by arrows with the increasing temperature and time increase along the direction of arrows. The labels near each unfolding event are the domain pairs that dissociate in the event. When several events occur close in time but not in a conserved temporal order, they are organized as one cluster without separations by arrows.

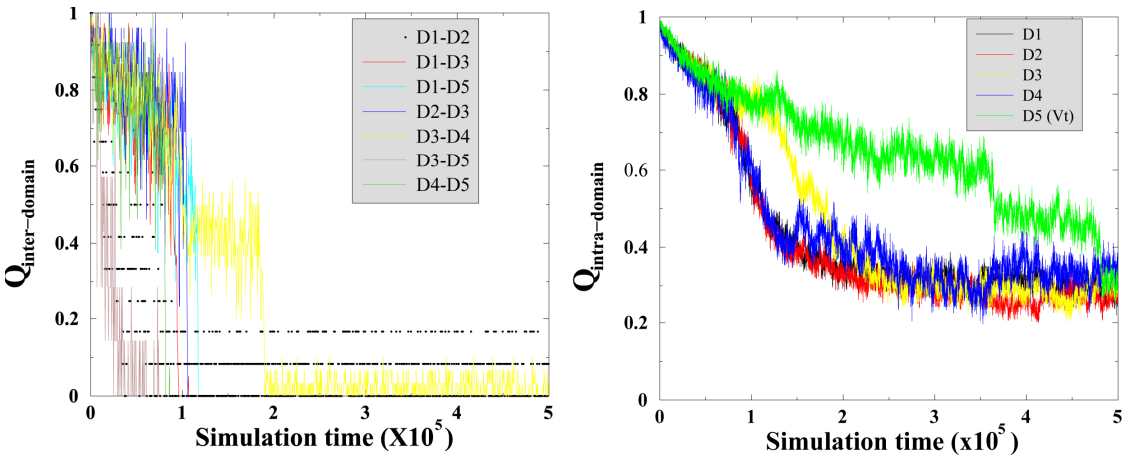

FIGURE 3.4. A sample trajectory from kinetic unfolding simulations. The fractions of native contacts *a,* between domains, and *b*, within domains are plotted as a function of simulation time.

The role of talin binding in vinculin activation

The binding of talin to vinculin was postulated as an important mechanism for vinculin activation (Bois et al., 2006; Izard and Vonrhein, 2004; Izard et al., 2004). It was proposed that the binding of talin to vinculin induces a conformational change of D1 domain of vinculin, which disrupts the interaction between D1 and Vt domains (Bois et al., 2006; Izard and Vonrhein, 2004; Izard et al., 2004). However, most supporting evidence for this postulate was not obtained in the context of the whole vinculin. To shed light on the role of talin binding in activation of the whole vinculin, we study the association of the vinculin binding site (VBS1, residue 605-636) from talin and the full-length vinculin. We model the interaction between vinculin and VBS1 by an effective potential based on the crystal structure of D1-VBS1 complex(Papagrigoriou et al., 2004) (Methods). We perform DMD simulation of the binding of this peptide to vinculin at various temperatures. We find that the disruption of D1-Vt interaction occurs in a cooperative manner as is observed for the ligand-

free state of vinculin. Interestingly, the temperature at which the D1-Vt interaction is completely disrupted is much lower in the presence of VBS1 ($T \approx 0.70$) than the ligand-free state of vinculin ($T \approx 0.76$), indicating that the binding of VBS1 peptide to vinculin significantly destabilizes the D1-Vt interaction (Figure 3.5 and the legend therein). On the other hand, the binding of VBS1 peptide to vinculin is insufficient to disrupt D1-Vt interaction under native conditions (Figure 3.5), suggesting a high free energy barrier between active and inactive state of vinculin in the presence of VBS1 in the simulation.

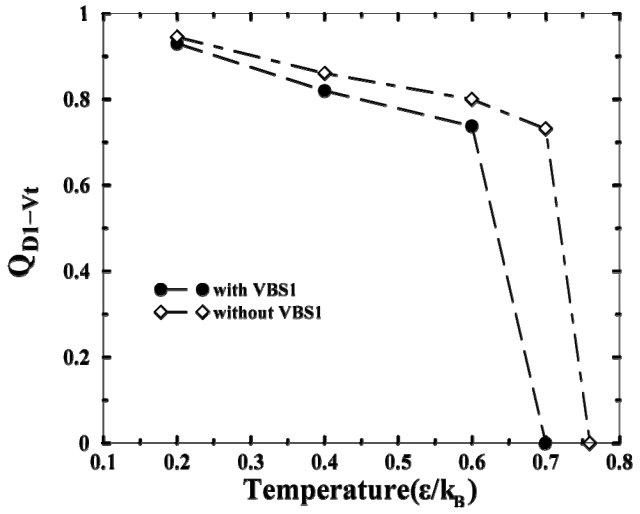

FIGURE 3.5. The fraction of native contacts $Q_{D1\text{-}Vt}$ made between D1 and Vt domains is plotted as a function of temperature. The Q value is calculated using trajectories from equilibrium simulations of the whole vinculin in the presence of VBS1 (605-636) peptide from talin. The temperature at which D1-Vt interaction is completely disrupted (Q=0) is around *T*=0.70, while in the absence of VBS1, this temperature is around *T*=0.76.

The influence of inherent flexibility on allostery of vinculin

The ability to adopt distinct conformations upon the binding of diverse ligands makes vinculin a multifarious allosteric protein. We find that the intrinsic flexibility of vinculin which is defined in the absence of ligand has a large influence on the allosteric properties of vinculin. The two most dominant motions characterized in our simulation in the absence of any ligand binding differ in details, but they both involve "opening" and "closing" motions between Vt domain and structural elements of the pincer-like structure formed by D1, D3 and D2. These motions represent the intrinsic tendencies of the transitions between the "closed", inactive state and an "open" state. The observed flexibility of vinculin is mainly determined by its native topology, suggesting that the native topology plays an important role in determining allosteric properties of vinculin.

Interestingly, it was found in recent studies(Tobi and Bahar, 2005) that the conformational changes of proteins upon ligand/protein binding are often well recapitulated by the intrinsic global motions of proteins in the unbounded state. These global motions are predominantly defined by native-contact topology of proteins(Tobi and Bahar, 2005). Hence, although the actual allosteric response is triggered by ligand binding, native topology may to a large extent define the plausible allostery of a protein.

The cooperative nature of vinculin activation makes it an allosteric switch

Despite the observation that the intrinsic plasticities of vinculin largely define the conformational changes associated with vinculin activation, the interaction interface between D1 and Vt remains rigid in all the dominant modes of principal motions near the native state.

Therefore, it is the ligand binding that is required for the disruption of D1-Vt interactions and vinculin acitivation. The current model of vinculin activation triggered by ligand binding is mainly based on two complementary lines of evidence. One comes from studies of the interaction between Vt and PtdIns(4,5)P_2(Bakolitsa et al., 1999; Gilmore and Burridge, 1996; Weekes et al., 1996). As suggested in the limited proteolysis experiments(Bakolitsa et al., 1999), the binding of PtdIns(4,5)P_2 to Vt triggers a pronounced conformational change of Vt that in turn inhibits the D1-Vt interactions and expose actin-binding and protein kinase C phosphorylation sites in Vt. The other line of evidence comes from more recent studies of the interaction between D1 and talin(Bass et al., 1999; Izard et al., 2004; Izard and Vonrhein, 2004). Remarkably, the structures of D1 bound to talin peptides show that the first half of D1 converts from a four- to a five-helix bundle upon binding, with a simultaneous repacking of the hydrophobic core(Izard et al., 2004). It was hypothesized that this "helix bundle conversion" produces a conformation that is incompatible with the D1–Vt interaction interface that can in turn disrupt the D1-Vt interactions. Instead of directly competing for the Vt binding surface of D1, talin alters the structure of D1. However, several other studies showed that larger fragments of talin bind poorly to full-length vinculin *in vitro*(Bakolitsa et al., 2004; Gilmore and Burridge, 1996). The addition of PtdIns(4,5)P2 is sufficient to increase the binding strength between talin and D1 fourfold(Gilmore and Burridge, 1996). Hence, the role of talin binding in the activation of full-vinculin activation remains controversial(DeMali, 2004).

We find that the binding of VBS1 peptide from talin to vinculin does significantly destabilize the D1-Vt interaction. Also, the VBS1 binding is insufficient to completely disrupt the D1-Vt interaction under native conditions. The latter observation is not fully

consistent with the recent experimental observations that the titration of VBSs from talin disrupts the D1-Vt interaction between the disjoined head and tail fragments of vinculin (complex formed between residue 1-840 and residue 879-1066) at room temperature(Bois et al., 2006). This result speaks to the possibility that our simulation may overestimate the free energy barrier between the active and inactive states of vinculin in the presence of VBS1. We speculate that the main factor leading to this overestimation is that the binding of VBS1 to the D1 domain in vinculin induces intermediate conformational changes stabilized by interactions that are absent in the final native D1-VBS1 structure, while the Gō potential used in our simulation (Methods) cannot capture those favorable interactions non-existent in the native structure. Therefore, a further characterization of these possible intermediate conformations in the kinetic experiments is of great interest.

We also observe a cooperative dissociation of between D1 and Vt in both thermodynamic and kinetic simulations with increasing temperature, suggesting that vinculin activation may not be a stepwise, but rather a cooperative, process. Once the binding of the ligands is able to overcome the activation barrier, vinculin rapidly becomes activated. The cooperative nature of vinculin activation may give rise to the switch-like property of vinculin in integrating the external signals (the specific ligand binding).

Recently, two studies pointed out that vinculin may play important regulatory roles rather than merely being a structural link between the actin cytoskeleton and adhesion receptors (DeMali et al., 2002; Subauste et al., 2004). First, it was found that the recruitment of the Arp2/3 complex to vinculin is required for efficient lamellipodial protrusion(DeMali et al., 2002), suggesting that vinculin might couple cell adhesion and membrane protrusion by localizing actin assembly to sites of newly engaged integrins(DeMali et al., 2002; DeMali

and Burridge, 2003). Second, vinculin regulates cell survival and migration by modulating the recruitment of paxillin to focal adhesions(Subauste et al., 2004). Therefore, the switch-like property and the requirement of multiple ligand binding for activation enables vinculin to function as an "AND" logic switch in rapid response to the regulatory signals and exert dynamic control over downstream events.

The synergistic contribution from domain-domain interactions to vinculin activation

The dissociations between Vt and D4, and Vt and D3 domains are two obligatory steps on the route to the disruption of the D1-Vt interaction in kinetic unfolding simulations. This observation suggests that besides the D1-Vt interaction, the direct interactions between Vt and other domains may also be essential in modulating vinculin activation. Consistent with our observation, Cohen et al.(Cohen et al., 2005) recently provided compelling biochemical evidence that the interactions between Vt and D4 are important in maintaining the auto-inhibited state of vinculin.

The dissociations between domains D1-Vt and the disassembly of the pincer-like structure (the dissociations between D2-D3, and D1-D3) occur in a cooperative manner (close to each other temporally) despite that the temporal order of the occurrence is not conserved. This observation raises a possibility that these domains may form a rigid cooperative unit, in which the significant dissociation between any pair of domains in this unit may subsequently trigger the dissociations between other domains upon even a slight external perturbation (such as ligand binding and temperature). Hence, we hypothesize that the disruption of D1-Vt interactions may involve a synergistic interplay between other domains. This hypothesis is experimentally-testable as it predicts that the mutations

weakening the interactions between D2 and D3, D1 and D3 may effectively destabilize the auto-inhibited state of vinculin.

CHAPTER 4

DETERMINING PROTEIN STRUCTURE USING INTER-RESIDUE PROXIMITY CONSTRAINTS

Introduction

Structure determination of protein is traditionally accomplished by X-ray crystallography or Nuclear Magnetic Resonance (NMR) based on inter-proton nuclear Overhauser enhancement (NOE). At the cost of months or years of laborious work, these methods can produce high-resolution structures at atomic-level. The structures of many proteins or protein complexes cannot be determined using these methods because of the specific physical and chemical properties of these proteins or complexes. On the other hand, the knowledge of these protein structures is very important to elucidate their biological functions due to the close relationship between structure and function. Therefore, there is pressing need for developing new methods to both increase the speed and broaden the target spectrum of protein structure determination with the rapid growth of the number of the identified proteins from genomic(Stoesser et al., 1999; Benson et al., 1999) and proteomic studies(Zhu et al., 2003) .

Emerging experimental methods, such as intra-molecular cross-linking(Cohen and Sternberg, 1980; Swaney, 1986) coupled with mass spectrometry (MS)(Fenn et al., 1989; McLafferty et al., 1999; Qin and Chait, 1996; Wang and Chait, 1994), fluorescence resonance energy transfer (FRET)(Dong et al., 2000; Cardullo and Parpura, 2003), electron paramagnetic resonance (EPR)(Hubbell et al., 1998; Lakshmi and Brudvig, 2001), and

paramagnetic relaxation enhancement (PRE)(Gillespie and Shortle, 1997b; Gillespie and Shortle, 1997a) allow for the determination of inter-residue proximity constraints, i.e. the typical distance separation range between two residues. These constraints have been used for fold identification (Albrecht et al., 2002; Young et al., 2000), as well as structure determination when other information is combined(Perozo et al., 1998; Gaponenko et al., 2000). With as few as 18 intra-molecular constraints derived from cross-linking and MS experiments, Young et al. (Young et al., 2000) identified the fold of the bovine basic fibroblast growth factor (FGF)-2 by selecting the structures consistent with the constraints from a pool of models generated by threading program. Using the constraints obtained by PRE (Gaponenko et al., 2000) and the known secondary structure constraints obtained from analysis of the nuclear Overhauser enhancement spectroscopy (NOSEY) data, the barnase structure was calculated with backbone root-mean-square-deviation (RMSD) less than 3Å from the crystal structure.

In practice, the number of inter-residue proximity constraints is often limited. Therefore, an important question is: what is the minimal number of inter-residue constraints needed to determine the fold of a protein? Several earlier studies (Aszodi et al., 1995; Smithbrown et al., 1993; Lund et al., 1996) addressed closely-related problems and offered significant insights into the relationship between protein structure and geometric constraints. However, *a priori* known secondary structures of protein were assumed in these studies and this assumption significantly limited their applications. More recent studies pioneered the combination of *de novo* structure prediction methods utilizing knowledge-based force field with limited NMR constraints to facilitate protein structure determination (Li et al., 2003; Meiler and Baker, 2005; Meiler and Baker, 2003; Skolnick et al., 1997). These studies exemplified that the

combination of a well-developed knowledge-based force field and experimental constraints is able to greatly improve the efficiency of structure determination. Here, we aim to offer a more general insight into the problem regarding minimal number of inter-residue constraints required for determining protein fold. First, we treat all distance constraints equally irrespective of the secondary or tertiary constraints and, therefore, have no *a priori* assumption about the constraints sets. Second, other than the distance constraints, bonded terms for chain connectivity, and steric exclusions between atoms, we do not apply any external force field when determining of protein structure, thereby ensuring the independence of the results on any specific force field. Therefore, we are able to obtain general insights into this problem with these minimal assumptions.

Another important question is how the distinct structural features of constraints differentiate their ability in determining the native fold and whether there is a rational strategy to select the inter-residue constraints that feature higher fidelity in structure determination. In such approaches as cross-linking/MS, FRET, EPR and PRE, one can in principle choose various sets of constraints by engineering the chemically-active residues into different positions in the protein. Thus, if such strategy for constraint selection exists, it can help decide what constraints best determine the protein structure.

We perform discrete molecular dynamics (DMD)(Dokholyan et al., 2003; Dokholyan et al., 1998; Smith et al., 1997; Zhou and Karplus, 1996) of eleven structurally-diverse protein domains subject to various sets of inter-residue proximity constraints and generate the corresponding structural ensembles. The typical inter-residue constraints obtained from experiments only contain the information of upper bound within which two atoms are distant from each other. In different types of experiments, this upper distance bound can vary

between ≈3 Å and >20Å. Within the upper distance bound, the exact distance between two atoms remains undetermined. Therefore the constraints with larger upper distance bound have larger uncertainties of inter-residue distances. Here, we do not consider the constraints with large upper distance bound (>10 Å), since such constraints are subject to larger uncertainties than more proximal constraints, and, consequently, without additional more precise structural information, these large-separation constraints are not useful in structure determination. For simplicity, we focus on the inter-residue constraints with a uniform upper distance bound 7.5 Å between C_β atoms (C_α for Gly). This upper distance bound is usually used to define the contact map of a protein structure, from which the native structure can be faithfully determined (Dokholyan et al., 2003). We study the dependencies of RMSD (from the native structure) of constructed structural ensembles on varying numbers of randomly selected constraints. We also attempt to identify rational strategies for selecting the constraints that result in higher fidelity in structure determination. A feasible rational strategy requires a quantitative measure that can distinguish the performance of constraints in structure determination. Thus, we calculate several topological properties of the selected constraints and test whether these properties dictate the performance of these constraints.

Material and methods

DMD simulation and four-bead protein model

We perform DMD(Dokholyan et al., 1998; Rapaport DC, 1997; Zhou and Karplus, 1996; Smith et al., 1997) simulations using a four-bead protein model(Ding et al., 2003), in which each residue is represented by three backbone beads N, C, C_α and one side-chain bead C_β (only C_α for Gly). The detailed implementation of the covalent bonds and constraints that

maintain the correct geometry of each residue and the peptide connectivity is described by Ding et al.(Ding et al., 2003).

Clustering methods

Clustering is the procedure that categorizes or groups similar entities together based on quantitative distance (similarity) measures. To perform clustering of structures in the present study, we define the distance between two structures as their mutual RMSD. The smaller RMSD from each other, the higher similarity there is between two structures. We perform hierarchical agglomerative clustering(Everitt et al., 2001) by first finding the two entities that have the minimal distance between them. Joining those two entities into a cluster, the method then searches for the minimal distance between two entities, but taking those entities that have already been clustered as a single unit. This process is repeated until there are no more entities to cluster. The process of the hierarchical agglomerative clustering is summarized in a tree-like diagram, called dendrogram(Everitt et al., 2001). The root of the dendrogram, which is at the top (zeroth) level, is one cluster containing all the entities (structures). The leaf nodes at the bottom level of the dendrogram correspond to isolated entities before clustering. From the top to the bottom level of the dendrogram, the number of emerging clusters increases. There are three different hierarchical agglomerative clustering methods applied: *single linkage*, *complete linkage* and *average linkage*. In *single linkage*, the minimal distance between members of the two clusters is taken as the cluster distance. In *complete linkage*, the maximal distance between members of clusters is taken. In *average linkage*, the average distance between members in the clusters is taken. All the clustering in this work is performed using the program OC(Barton, 2002).

Selection of constraints from contact map

A protein contact map is a set of contacts defined as follows: if the distance between C_β atoms (C_α for Gly) from two residues *i* and *j* in the native structure is within 7.5Å, residues *i* and *j* are considered to form a native contact. Here, an inter-residue proximity constraint corresponds to a native contact in the contact map. We will use native contacts and constraints interchangeably hereafter. For each native contact between residues *i* and *j*, the contact distance is defined as $|i-j|$. We exclude the constraints with the contact distance $|i-j| \leq 2$, since these constraints are mainly defined by polypeptide connectivity.

The contact order(Plaxco et al., 1998) of a set of constraints is defined as the average $|i-j|$ taken over all constraints in the set. In simulations, we use different strategies for selecting constraints from the contact map. In the *random selection* procedure, the constraints are randomly selected from the contact map. In the *contact order-ranked* (COR) method, the constraints are selected from the contact map sequentially according to the descending order of the corresponding $|i-j|$ values ($i,j=1,2,\ldots N$) .

Generation of coarse-grained structural ensembles satisfying a given set of constraints

We incorporate a given set of inter-residue constraints into the simulations as effective interactions between C_β atoms (C_α for Gly) of corresponding residues:

$$U_{ij} = \begin{cases} +\infty, & |r_i - r_j| \leq a_0 \\ -\Delta_{ij}, & a_0 < |r_i - r_j| \leq a_1 \\ 0, & |r_i - r_j| > a_1 \end{cases}, \quad (4.1)$$

where a_0 is the hard core diameter, a_1 is the upper-bound of distance constraints which is 7.5Å. All $\mathbf{r}_i$ and $\mathbf{r}_j$ are the Cartesian coordinates of C_β atom (C_α for Gly) of *i*th and *j*th residue respectively. $||\Delta_{ij}||$ is a matrix with elements $\Delta_{ij}=1$ if there is a constraint between residue *i*

and j, Δ_{ij}=0 otherwise. For each set of constraints, prior to the production simulations, we perform simulations from the fully-extended protein state at temperature T=2.0 to T=0.1 (in units of the inverse Boltzmann constant k_B^{-1}). Then, we perform production simulations at T=0.1 for 10^5 time units. We choose five trajectories starting from different initial conditions in which all given constraints are satisfied and there is no "mirror structure"– a structure that satisfies all given constraints, but the transformation that superimposes it with the native structure is an improper rotation (Kabsch, 1976). The trajectories of mirror structures are automatically eliminated by clustering the structures from trajectories that satisfy all the given constraints (Clustering methods). Regardless of the different clustering methods applied, the "mirror structures" always appear as a distinct cluster at the first level of dendrograms(Everitt et al., 2001) (Clustering methods). In addition, the "mirror structures" tend to have higher energies than other structures that satisfy all constraints after all-atom reconstruction. Therefore, we can recognize "mirror structures" without *a priori* knowledge of the native structure.

Topological properties of constraints in the contact map graph

We construct the contact map graph by representing each residue as a node and each constraint in the contact map as an edge. Each set of constraints represents a sub-graph of the contact map graph. There are four topological properties of constraint sets in the graph inspected in the current study.

The *shortest path*(Bollobás, 1998) between two nodes i and j in the network are defined as the paths traversing minimal number of edges among all paths connecting i and j. The *shortest path length* between two nodes is the number of edges traversed by the shortest

paths. The *shortest path length* of a set of constraints is the average of *shortest path lengths* between all nodes in the set.

The *betweennes centrality* $C_B(l)$ of an edge *l* in the network is defined as follows(Freeman, 1977):

$$C_B(l) = \sum_{s \neq t \in V} \frac{\sigma_{st}(l)}{\sigma_{st}}, \quad (4.2)$$

where σ_{st} is the total number of shortest paths connecting node *s* and *t*; $\sigma_{st}(l)$ is the number of shortest paths connecting node *s* and *t* that pass through edge *l.* The *betweenness centrality* of a set of constraints is the average of *betweenness centrality* taken over all edges in the set.

The *clustering coefficient* of a node(Watts and Strogatz, 1998; Wasserman and Faust, 1994) *i* in the network is defined as the ratio between the number E_i of edges that actually exist between all k_i nodes directly connected to node *i*, and the maximal possible number $k_i(k_i - 1)/2$ of edges between these k_i nodes:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}. \quad (4.3)$$

The *clustering coefficient* of a set of constraints is the average of *clustering coefficients* taken over all nodes in the set.

The *degree* of a node *i*(Bollobás, 1998) in the network is defined as total number of nodes having edges directly connected to it. The *degree* of a set of constraints is the average of *degrees* taken over all nodes in the set.

Results and Discussion

We study nine protein domains chosen from the CATH protein structure classification database(Orengo et al., 1997). The CATH database is a hierarchical domain classification of

protein structures in the PDB. There are four major "top to bottom" levels of classification of protein structures in the CATH database, which are Class, Architecture, Topology and Homologous superfamilies. For the broad coverage of the structural space of all proteins, we choose nine protein domains representing all categories: mainly-α, mainly-β, α-β, and few secondary structures (Table 4.1) at the top level Class in CATH. To minimize the length dependence of the result, all protein domains are chosen to be approximately 60 residues long. For each protein domain, we first determine the contact map based on its native structure (Methods). The contact map contains all inter-residue proximity constraints with a cutoff distance of 7.5 Å between C_β atoms (C_α for Gly). For each fraction of constraints, 10%, 30%, 50%, 70%, 90%, we then generate six constraint sets with the corresponding number of constraints selected from the contact map. One set is generated by the COR method and the other five are by a random selection (Methods). Then, we perform DMD simulations of the simplified protein model (Methods) to generate the coarse-grained conformational ensembles satisfying these constraint sets.

Length	PDB code	CATH code	Class(C)
60	1GO3(48:107)	1.10.150.80	Mainly-α
49	1DD3(1 : 49)	1.20.58.20	Mainly-α
62	1NXB(1 : 62)	2.10.60.10	Mainly-β
60	1VIE(19 : 78)	2.30.30.60	Mainly-β
60	1JDC(358:417)	2.60.40.1180	Mainly-β
61	1IGD(1 : 61)	3.10.20.10	α- β
60	1BXY(1 : 60)	3.30.70.700	α- β
65	1E4F(237:301)	3.90.640.10	α- β
60	1D0D (1 : 60)	4.10.410.10	Few secondary-structures

TABLE 4.1. Nine protein domains

Approximately 70% of all proximity constraints derived from the native structures are sufficient to determine the protein folds within an average RMSD of 3.4 Å

We determine the average and standard deviation of RMSD of the structural ensembles, subject to five randomly-selected constraint sets, as the function of the fraction of constraints applied (Figure 4.1). For each fraction of constraints, 10%, 30%, 50%, 70%, and 90% in Fig. 4.1, we only show the largest and the smallest average RMSD with their corresponding standard deviations. We observe a sharp decrease of the average RMSD as the fraction of constraints increases. Although the exact fraction at which this cooperative transition occurs varies for different domains, this transition typically occurs at less than ≈30% fraction of constraints. For the protein domains under study, the 30% fraction of constraints correspond to 0.65 ± 0.05 constraints per residue (the number of constraints applied per length of a domain). In addition, we find that 70% fraction of all proximity constraints derived from the native structure are sufficient to determine the fold of a domain with an average RMSD of ≤3.4 Å. The 70% fraction of constraints corresponds to ~90 constraints for a 60 residue protein and 1.51 ± 0.11 constraints per residue. Noticeably, for different folds, the minimal fraction of constraints required to determine the native fold varies significantly (Figure 4.1), reflecting unique topological characteristics of distinct folds. For example, for protein domains 1GO3 (mainly-alpha), 1NXB (mainly-beta), 1VIE (mainly-beta), 1BXY (alpha-beta) and 1E4F (alpha-beta), 50% fraction constraints are sufficient to determine the structural ensembles within an average RMSD of 4 Å from the native structures, while it is not the case for other domains. Therefore 70% is only a conservative estimate.

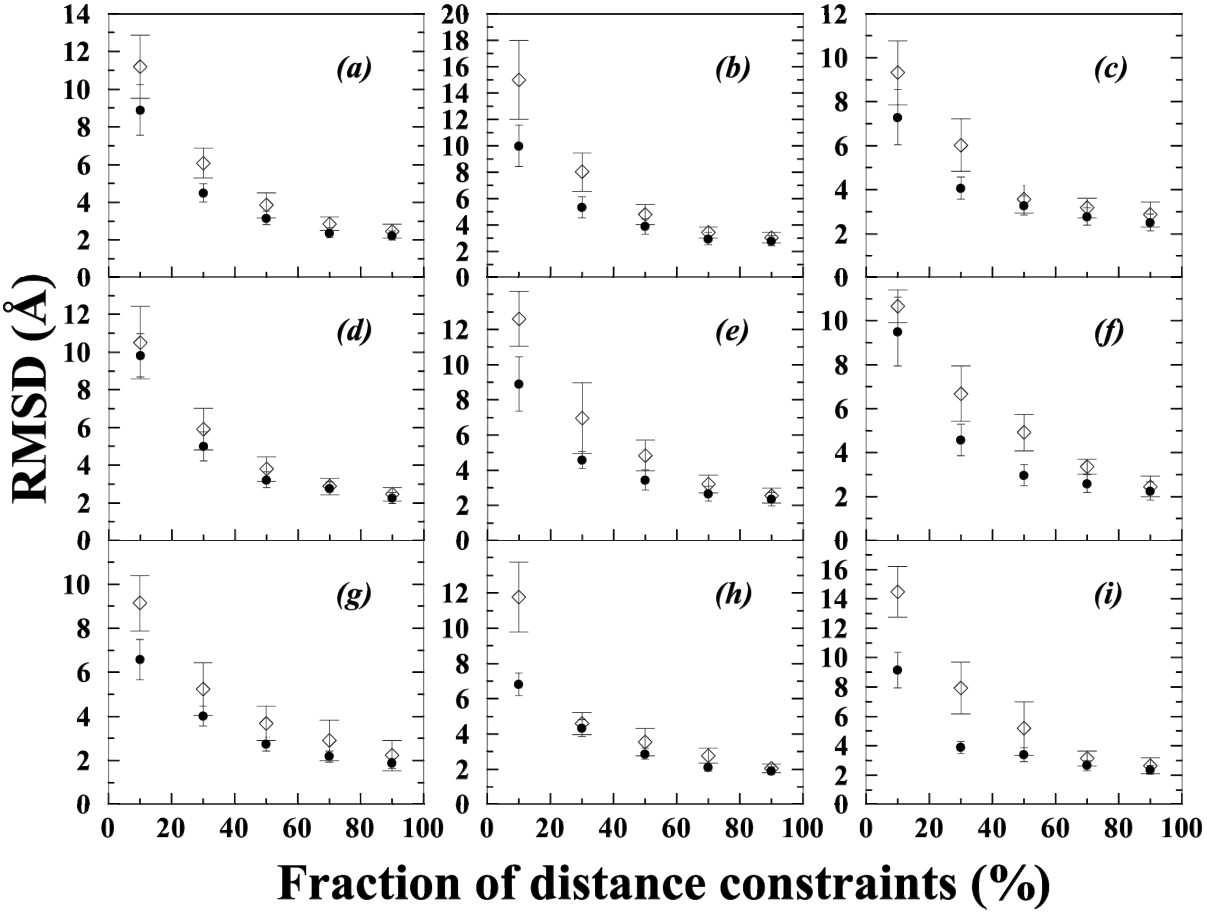

FIGURE 4.1. The average and the standard deviation of RMSD of the structural ensembles subject to 10%, 30%, 50%, 70%, 90% randomly-selected constraints for nine domains: (a) 1GO3 (48-107), (b) 1DD3 (1-49), (c) 1NXB (1-62), (d) 1VIE (19-78), (e) 1JDC (358-417), (f) 1IGD (1-61), (g) 1BXY (1-60), (h) 1E4F (237-301), (i) 1D0D (1-60). For each fraction of constraints, we only show the largest (◊) and the smallest (●) average RMSD and their corresponding standard deviations (out of five ensembles).

According to the theoretical work by Reva et al. (Reva et al., 1998) and others(Ding et al., 2005), the RMSD distribution for a ~60-residue protein with randomly selected/constructed globular protein-like structure is Gaussian with the average value of 11 Å and the standard deviation of 2 Å. Therefore the probability of observing a structure within 3.4 Å RMSD from

the native structure by chance is less than $7x10^{-5}$. It is statistically significant to conclude that a structural ensemble satisfying 70% native contacts is close to the native structure.

The current study focuses on the relatively small domains as test cases. For larger domains, the total number of the constraints required for determining the native fold will increase accordingly. To see whether the current results can be extrapolated to larger domains, we study two other domains with the length of approximately 105 residues (Table 4.2). We find that for these two domains, 70% randomly-selected native contacts are also sufficient to determine the native folds (Figure 4.2). Therefore, we expect that our results will also hold for larger single-domain proteins. However, it remains to uncover in future studies to what extent our result can be extended to the case of multi-domain proteins.

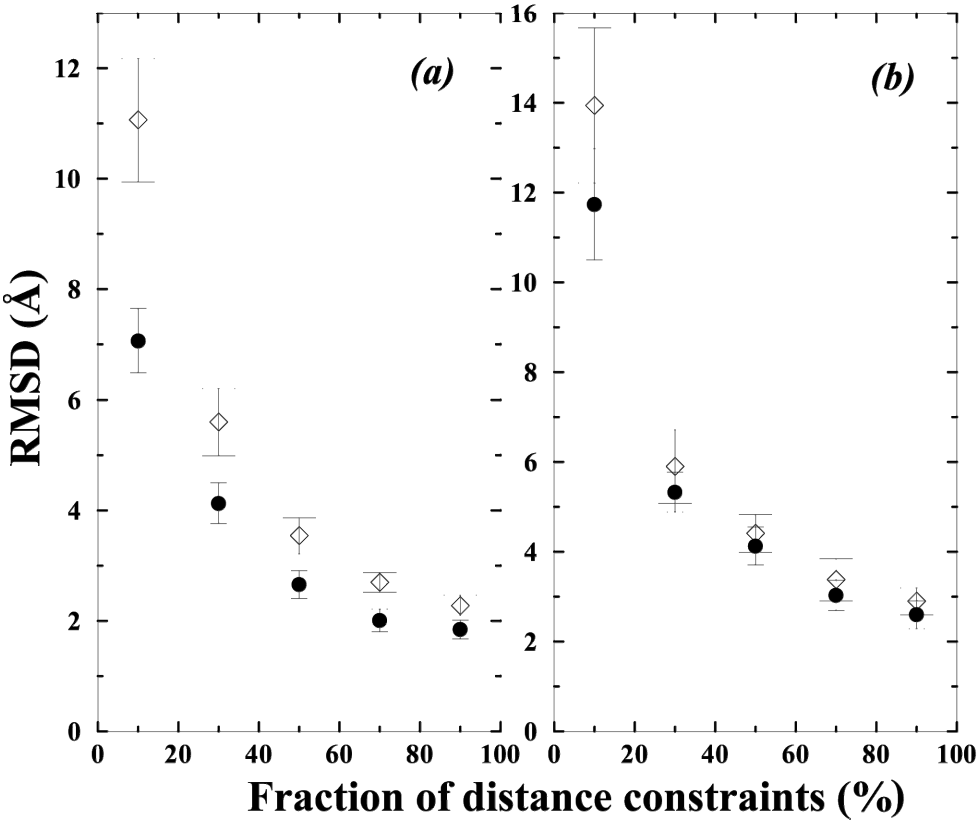

FIGURE 4.2. The average and the standard deviation of RMSD of the structural ensembles subject to 10%, 30%, 50%, 70%, 90% randomly-selected constraints for (a) 1FMT (209-313)

and (b) 1K79 (333-436). For each fraction of constraints, we only show the largest (◊) and the smallest (●) average RMSD and their corresponding standard deviations (out of five ensembles).

Length	PDB code	CATH code	Class(C)
104	1K79(333:436)	1.10.10.10	Mainly-α
105	1FMT(209:313)	3.10.25.10	α-β

Table 4.2. Two large protein domains

Random constraint selection often outperforms .rational contact order-based selection strategy

For each fraction of constraints, 10%, 30%, 50%, 70%, and 90%, we find that the structural ensembles subject to six constraint sets (five randomly-selected sets and one selected by the COR method) have distinct RMSD from the native structure. For example, in the case of the protein domain 1E4F (237-301) (Fig. 4.1h), the structural ensembles, subject to various constraint sets utilizing 10% of the constraints, have an average RMSD ranging from 7 – 12 Å. In 1E4F (237-301) 10% of the constraints corresponds to 13 constraints. These results indicate that using a sparse number of constraints (~13), it is possible to reconstruct a structural ensemble with an average RMSD of 7 Å. When other information, such as homology or secondary structure information, is incorporated, it is not surprising that the structure ensembles can be determined with higher accuracy(Young et al., 2000; Gaponenko et al., 2000). The large performance variation of different constraint sets also suggests that proper selections of constraints can significantly improve the efficiency of structure determination.

The constraint sets that have large values of contact order (i.e. have more long-range constraints) offer more information about global than local structural properties.

Correspondingly, the constraint sets that have small values of the contact order offer more information about local than global structural protein properties. Intuitively, global structural information plays major role in determining protein folds. Hence, we speculate that the constraint sets containing more long-range constraints lead to structural ensembles with smaller RMSD from the native structure. Since the separation of residues in the constraint sets can be quantified by contact order (Methods), we develop a rational strategy, COR method (Methods), for constraint selection to predominantly favor long-range constraints. Surprisingly, we find that the conformational ensembles, corresponding to the constraint sets selected by the COR method, often have larger RMSD than those ensembles corresponding to the constraint sets selected randomly (Fig. 4.3). Clearly, randomly-selected sets have a broader distribution of contact distance than the sets selected by the COR method (Fig. 4.4), which is due to the fact that random selection tends to include a mixture of short-range and long-range constraints, while the COR method favor long-range constraints. The observed difference of the performance between random selection and the COR method suggests that accurate structure determination depends on a blend of global and local structural information rather than global structural information alone.

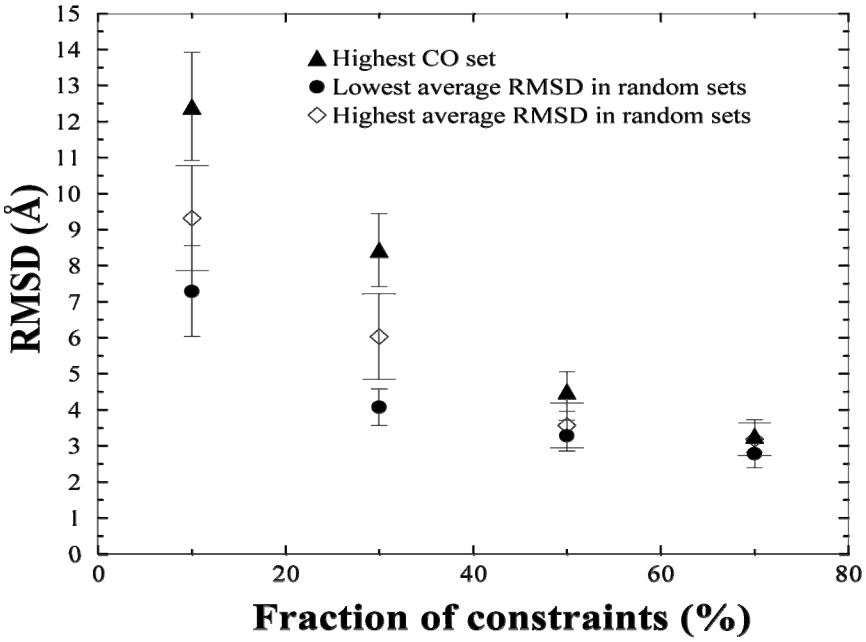

FIGURE 4.3. Neurotoxin b (PDB code: 1NXB). The average and standard deviation of RMSD of the structural ensembles subject to 10%, 30%, 50%, 70%, 90% constraints selected by the contact order-ranked method and random strategy: Triangle (▲), the structural ensembles subject to constraint sets selected by the contact order-ranked method (CO); Diamond (◊), the highest average RMSD and corresponding standard deviation of the structural ensembles subject to randomly-selected constraint sets; Filled circle (●), the lowest average RMSD and corresponding standard deviation of the structural ensembles subject to randomly-selected constraint sets.

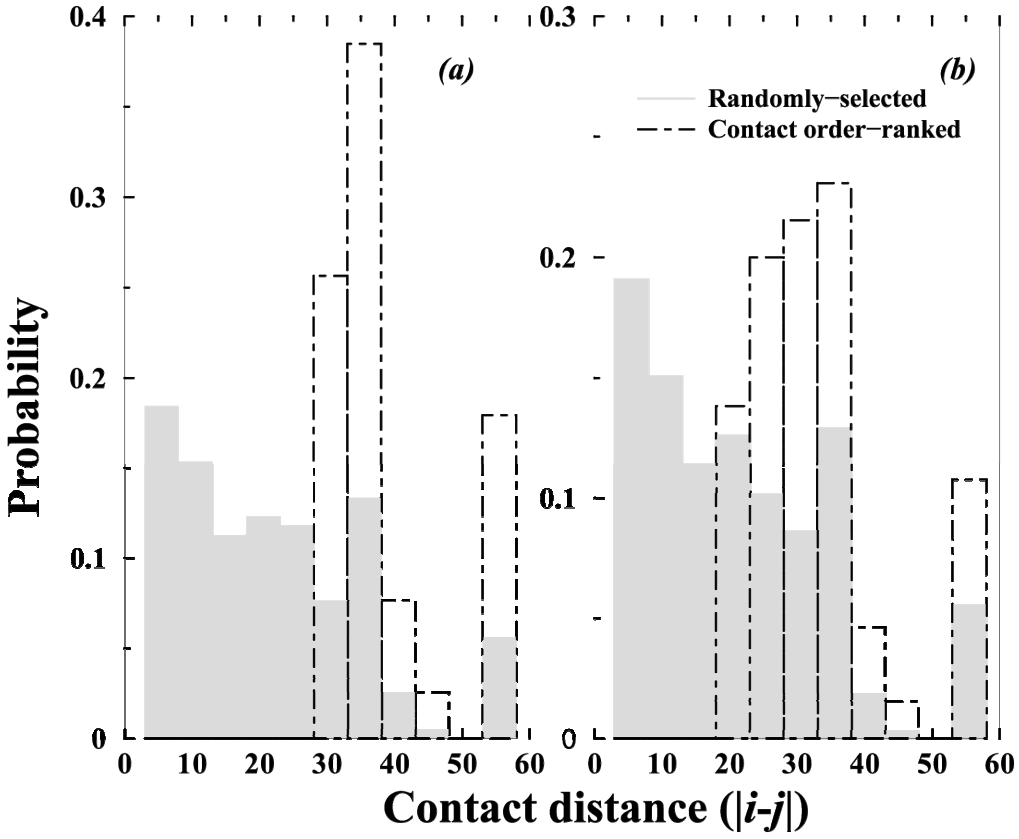

FIGURE 4.4. Neurotoxin b (PDB code: 1NXB). The histograms of contact distance (|*i-j*|) for: (a) 30% constraints selected by random strategy (grey) and the contact order-ranked method (dot-dashed); (b) 50% constraints selected by random strategy (grey) and the contact order-ranked method (dot-dashed).

Simple topological properties of the selected constraints do not correlate with the performance of these constraints in structure determination

To explore other potential factors of structure determination performance, we examine several topological properties of constraint sets in the *contact map graph* (Methods), where each residue is represented as a node and each constraint is represented as an edge. The spatial relationships between the residues in the protein structure are mapped to the connections between the nodes in the contact map graph. Each set of selected constraints represents a sub-graph of the contact map graph. First, we study the correlation between the

betweenness centrality (Methods) of a constraint set and the performance of this set in structure determination. The betweenness centrality of an edge measures how often this edge lies on the shortest path between all node pairs in the graph and indicates its significance in the connectivity in the whole graph(Freeman, 1977). The higher betweenness centrality an edge has, the more important it is in connecting the whole graph. We expect that the constraints with larger betweenness centrality contain more structural information on the spatial connection between different elements of the protein structure and, thus, have more prominent effects on protein structure determination. Although the betweenness centrality is a sensitive measure of network topology(Dokholyan, 2005), we do not observe correlation between the betweenness centrality and the performance of a constraint set (Table 4.3). For example, in the case of domain 1NXB(1-62) a constraint set with a larger average betweenness centrality leads to an ensemble with smaller average RMSD, while in the case of domain 1IGD(1-61) (Table 4.3) a constraint set with a larger average betweenness centrality leads to an ensemble with larger average RMSD.

Constraint sets (random)	1NX8 (10%)	1IGD (10%)	1NX8 (30%)	1IGD (30%)
Largest average RMSD	18.86	26.58	24.50	24.72
Smallest average RMSD	23.37	19.46	26.18	22.04

TABLE 4.3. The differences in the edge betweenness centrality between the constraint sets corresponding to the ensembles with the largest average RMSD, and the constraint sets corresponding to the ensembles with the smallest average RMSD among the randomly-selected set. For illustration, we only show the result of the constraint sets contain 10% and 30% constraints for two domains (1NXB (1-62), 1IGD (1-61)), respectively.

Next, we study the correlation between the performance of a given constraint set and other topological properties, such as the node degree, the clustering coefficient, and the shortest path length (Methods). These three topological properties describe various global and local features of networks (graphs). The degree of a graph node is the number of nodes directly connected to it by edges (i.e. the number of neighbors). The nodes with higher degrees correspond to the highly interacting residues in the native structure. The clustering coefficient of a node is the probability of its neighboring nodes being directly connected. When the clustering coefficient of a node is closer to one, the corresponding residues of its neighboring nodes have higher probability to form contacts. A smaller shortest path length between two nodes corresponds to a shorter communication path between the corresponding residues mediated by amino acid contacts in the native structure. These three topological properties have been used to characterize the specific features of various networks(Albert and Barabasi, 2002). Especially, it was shown that the topological properties, such as betweenness centrality or shortest path length, can be used to identify the key residues in stabilizing the folding transition state of proteins, as well as to differentiate the pre- and post-transition state of proteins along folding pathways(Dokholyan et al., 2002; Vendruscolo et al., 2002). However, similar to the case of betweenness centrality, we do not observe correlation between any of these topological properties and the performance of constraint sets (data not shown).

These findings indicate that the RMSD of a conformation to the native structure is a multi-variable function of different topological properties of constraints. Therefore the accurate structure determination requires a combination of constraints with composite structural features, which are not characterized by any single topological property.

Conclusions

Without *a priori* knowledge of secondary structures of proteins and the application of a knowledge-based/physical force field in simulations, we have been able to determine the fold of domains (with an average RMSD of ≤3.4 Å) for eleven structurally diverse protein domains using approximately 70% of all proximity constraints derived from their native structures. This finding by no means offers a comprehensive answer to the question-what is the minimal number of inter-residue proximity constraints needed to determine the fold of a protein. Rather, it offers a theoretical estimation of the lower-bound of this minimal number. We believe that this estimation is an important starting point for further studies. It is important to note, that in the current study is that we do not consider any experimental errors of constraints in the simulations. It is expected that different types of experimental errors can have distinct effects on the determined protein structure and these effects deserve further studies.

In addition, we find that randomly-selected constraints often outperform the constraints representing global structural features, suggesting that both local and global structural features are important in determining the fold of a protein. Further, we do not observe any correlation between various topological properties of the selected constraints, emphasizing different structural features, and the performance of these constraints. Both these findings indicate that a rational strategy based on a quantitative measure that can distinguish the performance of constraints in structure determination is yet to be developed. Due to the significant complexity inherent in the mapping from constraints to the protein structures, more work is needed to understand how to build a minimum set of structure determining constraints.

REFERENCES

Abe,H. and Go,N. (1981). Non-Interacting Local-Structure Model of Folding and Unfolding Transition in Globular-Proteins .2. Application to Two-Dimensional Lattice Proteins. Biopolymers *20*, 1013-1031.

Albert,R. and Barabasi,A.L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics *74*, 47-97.

Albrecht,M., Hanisch,D., Zimmer,R., and Lengauer,T. (2002). Improving fold recognition of protein threading by experimental distance constraints. In Silico. Biol. *2*, 325-337.

Amadei,A., Linssen,A.B., and Berendsen,H.J. (1993). Essential dynamics of proteins. Proteins *17*, 412-425.

Arold,S.T., Hoellerer,M.K., and Noble,M.E. (2002). The structural basis of localization and signaling by the focal adhesion targeting domain. Structure *10*, 319-327.

Aszodi,A., Gradwell,M.J., and Taylor,W.R. (1995). Global Fold Determination from A Small Number of Distance Restraints. Journal of Molecular Biology *251*, 308-326.

Bai,Y., Milne,J.S., Mayne,L., and Englander,S.W. (1993). Primary structure effects on peptide group hydrogen exchange. Proteins *17*, 75-86.

Bai,Y., Milne,J.S., Mayne,L., and Englander,S.W. (1994). Protein stability parameters measured by hydrogen exchange. Proteins *20*, 4-14.

Bai,Y., Sosnick,T.R., Mayne,L., and Englander,S.W. (1995). Protein folding intermediates: native-state hydrogen exchange. Science *269*, 192-197.

Bakolitsa,C., Cohen,D.M., Bankston,L.A., Bobkov,A.A., Cadwell,G.W., Jennings,L., Critchley,D.R., Craig,S.W., and Liddington,R.C. (2004). Structural basis for vinculin activation at sites of cell adhesion. Nature *430*, 583-586.

Bakolitsa,C., de Pereda,J.M., Bagshaw,C.R., Critchley,D.R., and Liddington,R.C. (1999). Crystal structure of the vinculin tail suggests a pathway for activation. Cell *99*, 603-613.

Bakolitsa,C., de Pereda,J.M., Bagshaw,C.R., Critchley,D.R., and Liddington,R.C. (1999). Crystal structure of the vinculin tail suggests a pathway for activation. Cell *99*, 603-613.

Barton,G.J. (2002). OC - A cluster analysis program. (Scotland, UK: University of Dundee).

Bass,M.D., Smith,B.J., Prigent,S.A., and Critchley,D.R. (1999). Talin contains three similar vinculin-binding sites predicted to form an amphipathic helix. Biochem. J. *341*, 257-263.

Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F.F., Rapp,B.A., and Wheeler,D.L. (1999). GenBank. Nucleic Acids Research *27*, 12-17.

Berendsen,H.J. and Hayward,S. (2000). Collective protein dynamics in relation to function. Curr. Opin. Struct. Biol. *10*, 165-169.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., and Bourne,P.E. (2000). The Protein Data Bank. Nucleic Acids Research *28*, 235-242.

Binder,K. (1995). The Monte Carlo method in condensed matter physics - Introduction. Monte Carlo Method in Condensed Matter Physics, Second, Corrected and Updated Edition *71*, 1-22.

Bois,P.R., O'Hara,B.P., Nietlispach,D., Kirkpatrick,J., and Izard,T. (2006). The vinculin binding sites of talin and alpha-actinin are sufficient to activate vinculin. J. Biol. Chem. *281*, 7228-7236.

Bollobás,B. (1998). *Modern Graph Theory*. (New York: Springer).

Borreguero,J.M., Dokholyan,N.V., Buldyrev,S.V., Shakhnovich,E.I., and Stanley,H.E. (2002). Thermodynamics and folding kinetics analysis of the SH3 domain form discrete molecular dynamics. J. Mol. Biol. *318*, 863-876.

Branden,C. and Tooze,J. (1999). Introduction to protein structure. (New York: Garland Pub).

Brown,N.R., Noble,M.E., Endicott,J.A., and Johnson,L.N. (1999). The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. Nat. Cell Biol. *1*, 438-443.

Cardullo,R.A. and Parpura,V. (2003). Fluorescence resonance energy transfer microscopy: theory and instrumentation. Methods Cell Biol. *72*, 415-430.

Cavanagh,J., Fairbrother,W.J., Palmer,A.G., and Skelton,N.J. (1996). Protein NMR spectroscopy : principles and practice. (San Diego: Academic Press).

Chamberlain,A.K., Handel,T.M., and Marqusee,S. (1996). Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH. Nat. Struct. Biol. *3*, 782-787.

Chen,H.C., Appeddu,P.A., Parsons,J.T., Hildebrand,J.D., Schaller,M.D., and Guan,J.L. (1995). Interaction of focal adhesion kinase with cytoskeletal protein talin. J. Biol. Chem. *270*, 16995-16999.

Clarke,J. and Fersht,A.R. (1996). An evaluation of the use of hydrogen exchange at equilibrium to probe intermediates on the protein folding pathway. Fold. Des *1*, 243-254.

Cohen,D.M., Chen,H., Johnson,R.P., Choudhury,B., and Craig,S.W. (2005). Two distinct head-tail interfaces cooperate to suppress activation of vinculin by talin. Journal of Biological Chemistry *280*, 17109-17117.

Cohen,F.E. and Sternberg,M.J. (1980). On the use of chemically derived distance constraints in the prediction of protein structure with myoglobin as an example. J. Mol. Biol. *137*, 9-22.

Cooley,M.A., Broome,J.M., Ohngemach,C., Romer,L.H., and Schaller,M.D. (2000). Paxillin binding is not the sole determinant of focal adhesion localization or dominant-negative activity of focal adhesion kinase/focal adhesion kinase-related nonkinase. Mol. Biol. Cell *11*, 3247-3263.

Critchley,D.R. (2000). Focal adhesions - the cytoskeletal connection. Curr. Opin. Cell Biol. *12*, 133-139.

DeMali,K.A. (2004). Vinculin--a dynamic regulator of cell adhesion. Trends Biochem. Sci. *29*, 565-567.

DeMali,K.A., Barlow,C.A., and Burridge,K. (2002). Recruitment of the Arp2/3 complex to vinculin: coupling membrane protrusion to matrix adhesion. J. Cell Biol. *159*, 881-891.

DeMali,K.A. and Burridge,K. (2003). Coupling membrane protrusion and cell adhesion. J. Cell Sci. *116*, 2389-2397.

Diestel,R. (2005). Graph theory. (Berlin: Springer).

Ding,F., Borreguero,J.M., Buldyrey,S.V., Stanley,H.E., and Dokholyan,N.V. (2003). Mechanism for the alpha-helix to beta-hairpin transition. Proteins-Structure Function and Genetics *53*, 220-228.

Ding,F., Buldyrev,S.V., and Dokholyan,N.V. (2005). Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. Biophysical Journal *88*, 147-155.

Ding,F., Dokholyan,N.V., Buldyrev,S.V., Stanley,H.E., and Shakhnovich,E.I. (2002a). Direct molecular dynamics observation of protein folding transition state ensemble. Biophysical Journal *83*, 3525-3532.

Ding,F., Dokholyan,N.V., Buldyrev,S.V., Stanley,H.E., and Shakhnovich,E.I. (2002b). Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. Journal of Molecular Biology *324*, 851-857.

Ding,F., Prutzman,K.C., Campbell,S.L., and Dokholyan,N.V. (2006). Topological determinants of protein domain swapping. Structure *14*, 5-14.

Dixon,R.D., Chen,Y., Ding,F., Khare,S.D., Prutzman,K.C., Schaller,M.D., Campbell,S.L., and Dokholyan,N.V. (2004). New insights into FAK signaling and localization based on detection of a FAT domain folding intermediate. Structure *12*, 2161-2171.

Dokholyan,N.V. (2005). The architecture of the protein domain universe. Gene *347*, 199-206.

Dokholyan,N.V., Borreguero,J.M., Buldyrev,S.V., Ding,F., Stanley,H.E., and Shakhnovich,E.I. (2003). Identifying importance of amino acids for protein folding from crystal structures. Methods Enzymol. *374*, 616-638.

Dokholyan,N.V., Buldyrev,S.V., Stanley,H.E., and Shakhnovich,E.I. (1998). Discrete molecular dynamics studies of the folding of a protein-like model. Fold. Des *3*, 577-587.

Dokholyan,N.V., Li,L., Ding,F., and Shakhnovich,E.I. (2002). Topological determinants of protein folding. Proceedings of the National Academy of Sciences of the United States of America *99*, 8637-8641.

Dong,W.J., Xing,J., Chandra,M., Solaro,J., and Cheung,H.C. (2000). Structural mapping of single cysteine mutants of cardiac troponin I. Proteins *41*, 438-447.

Englander,S.W. (2000). Protein folding intermediates and pathways studied by hydrogen exchange. Annu. Rev Biophys. Biomol. Struct. *29*, 213-238.

Englander,S.W., Mayne,L., Bai,Y., and Sosnick,T.R. (1997). Hydrogen exchange: the modern legacy of Linderstrom-Lang. Protein Sci. *6*, 1101-1109.

Englander,S.W., Sosnick,T.R., Englander,J.J., and Mayne,L. (1996). Mechanisms and uses of hydrogen exchange. Curr. Opin. Struct. Biol. *6*, 18-23.

Everitt,B.S., Landau,S., and Leese,M. (2001). Cluster analysis. (Oxford: Oxford University Press).

Fenn,J.B., Mann,M., Meng,C.K., Wong,S.F., and Whitehouse,C.M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science *246*, 64-71.

Fernandez,J.L.R., Geiger,B., Salomon,D., and Benzeev,A. (1993). Suppression of Vinculin Expression by Antisense Transfection Confers Changes in Cell Morphology, Motility, and Anchorage-Dependent Growth of 3T3-Cells. Journal of Cell Biology *122*, 1285-1294.

Fernandez,J.L.R., Geiger,B., Salomon,D., Sabanay,I., Zoller,M., and Benzeev,A. (1992). Suppression of Tumorigenicity in Transformed-Cells After Transfection with Vinculin Cdna. Journal of Cell Biology *119*, 427-438.

Freeman,L. (1977). A set of measures of centrality based upon betweeness. Sociometry *40*, 35-41.

Friedl,P. and Wolf,K. (2003). Tumour-cell invasion and migration: Diversity and escape mechanisms. Nature Reviews Cancer *3*, 362-374.

Fuentes,E.J. and Wand,A.J. (1998). Local dynamics and stability of apocytochrome b562 examined by hydrogen exchange. Biochemistry *37*, 3687-3698.

Gao,G.H., Prutzman,K.C., King,M.L., Scheswohl,D.M., Derose,E.F., London,R.E., Schaller,M.D., and Campbell,S.L. (2004). NMR solution structure of the focal adhesion targeting domain of focal adhesion kinase in complex with a paxillin LD peptide - Evidence for a two-site binding model. Journal of Biological Chemistry *279*, 8441-8451.

Gaponenko,V., Howarth,J.W., Columbus,L., Gasmi-Seabrook,G., Yuan,J., Hubbell,W.L., and Rosevear,P.R. (2000). Protein global fold determination using site-directed spin and isotope labeling. Protein Sci. *9*, 302-309.

Gillespie,J.R. and Shortle,D. (1997a). Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. J. Mol. Biol. *268*, 158-169.

Gillespie,J.R. and Shortle,D. (1997b). Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. J. Mol. Biol. *268*, 170-184.

Gilmore,A.P. and Burridge,K. (1996). Regulation of vinculin binding to talin and actin by phosphatidyl-inositol-4-5-bisphosphate. Nature *381*, 531-535.

Go,N. and Abe,H. (1981). Non-Interacting Local-Structure Model of Folding and Unfolding Transition in Globular-Proteins .1. Formulation. Biopolymers *20*, 991-1011.

Hanahan,D. and Weinberg,R.A. (2000). The hallmarks of cancer. Cell *100*, 57-70.

Hayashi,I., Vuori,K., and Liddington,R.C. (2002). The focal adhesion targeting (FAT) region of focal adhesion kinase is a four-helix bundle that binds paxillin. Nature Structural Biology *9*, 101-106.

Hildebrand,J.D., Schaller,M.D., and Parsons,J.T. (1995). Paxillin, A Tyrosine-Phosphorylated Focal Adhesion-Associated Protein Binds to the Carboxyl-Terminal Domain of Focal Adhesion Kinase. Molecular Biology of the Cell *6*, 637-647.

Hoellerer,M.K., Noble,M.E.M., Labesse,G., Campbell,I.D., Werner,J.M., and Arold,S.T. (2003). Molecular recognition of paxillin LD motifs by the focal adhesion targeting domain. Structure *11*, 1207-1217.

Hoffmann,E.D. and Stroobant,V. (2001). Mass Spectrometry: Principles and Applications. (New York: Wiley).

Hubbard,S.R. (1997). Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. Embo Journal *16*, 5572-5581.

Hubbell,W.L., Gross,A., Langen,R., and Lietzow,M.A. (1998). Recent advances in site-directed spin labeling of proteins. Curr. Opin. Struct. Biol. *8*, 649-656.

Hwang,T.L., Mori,S., Shaka,A.J., and vanZijl,P.C.M. (1997). Application of phase-modulated CLEAN chemical EXchange spectroscopy (CLEANEX-PM) to detect water-protein proton exchange and intermolecular NOEs. Journal of the American Chemical Society *119*, 6203-6204.

Hwang,T.L., van Zijl,P.C.M., and Mori,S. (1998). Accurate quantitation of water-amide proton exchange rates using the Phase-Modulated CLEAN chemical EXchange (CLEANEX-PM) approach with a Fast-HSQC (FHSQC) detection scheme. Journal of Biomolecular Nmr *11*, 221-226.

Ichiye,T. and Karplus,M. (1991). Collective Motions in Proteins - A Covariance Analysis of Atomic Fluctuations in Molecular-Dynamics and Normal Mode Simulations. Proteins-Structure Function and Genetics *11*, 205-217.

Izard,T., Evans,G., Borgon,R.A., Rush,C.L., Bricogne,G., and Bois,P.R.J. (2004). Vinculin activation by talin through helical bundle conversion. Nature *427*, 171-175.

Izard,T. and Vonrhein,C. (2004). Structural basis for amplifying vinculin activation by talin. Journal of Biological Chemistry *279*, 27667-27678.

Johnson,R.P. and Craig,S.W. (1995a). F-actin binding site masked by the intramolecular association of vinculin head and tail domains. Nature *373*, 261-264.

Johnson,R.P. and Craig,S.W. (1995b). The carboxy-terminal tail domain of vinculin contains a cryptic binding site for acidic phospholipids. Biochem. Biophys. Res. Commun. *210*, 159-164.

Jolliffe,I.T. (2002). Principal Component Analysis. (New York: Springer-Verlag).

Kabsch,W. (1976). A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A Crystal Physics Diffraction Theoretical and General Crystallography *5*, 922-923.

Katz,B.Z., Romer,L., Miyamoto,S., Volberg,T., Matsumoto,K., Cukierman,E., Geiger,B., and Yamada,K.M. (2003). Targeting membrane-localized focal adhesion kinase to focal adhesions - Roles of tyrosine phosphorylation and Src family kinases. Journal of Biological Chemistry *278*, 29115-29120.

Knight,B., Laukaitis,C., Akhtar,N., Hotchin,N.A., Edlund,M., and Horwitz,A.R. (2000). Visualizing muscle cell migration in situ. Current Biology *10*, 576-585.

Krishna,M.M., Hoang,L., Lin,Y., and Englander,S.W. (2004). Hydrogen exchange methods to study protein folding. Methods *34*, 51-64.

Kuriyan,J. and Cowburn,D. (1997). Modular peptide recognition domains in eukaryotic signaling. Annual Review of Biophysics and Biomolecular Structure *26*, 259-288.

Lakshmi,K.V. and Brudvig,G.W. (2001). Pulsed electron paramagnetic resonance methods for macromolecular structure determination. Curr. Opin. Struct. Biol. *11*, 523-531.

Lauffenburger,D.A. and Horwitz,A.F. (1996). Cell migration: A physically integrated molecular process. Cell *84*, 359-369.

Li,W., Zhang,Y., Kihara,D., Huang,Y.P.J., Zheng,D.Y., Montelione,G.T., Kolinski,A., and Skolnick,J. (2003). TOUCHSTONEX: Protein structure prediction with sparse NMR data. Proteins-Structure Function and Genetics *53*, 290-306.

Liu,G.H., Guibao,C.D., and Zheng,J. (2002). Structural insight into the mechanisms of targeting and signaling of focal adhesion kinase. Molecular and Cellular Biology *22*, 2751-2760.

Liu,J.S. (2001). Monte Carlo strategies in scientific computing. (New York: Springer).

Lo,S.H. (2006). Focal adhesions: What's new inside. Developmental Biology *294*, 280-291.

Lodish,H., Berk,A., Zipursky,L.S., Matsudaira,P., Baltimore,D., and Darnell,J. (2004). Molecular Cell Biology. (New York: W. H. Freeman).

Lund,O., Hansen,J., Brunak,S., and Bohr,J. (1996). Relationship between protein structure and geometrical constraints. Protein Science *5*, 2217-2225.

Maity,H., Lim,W.K., Rumbley,J.N., and Englander,S.W. (2003). Protein hydrogen exchange mechanism: Local fluctuations. Protein Science *12*, 153-160.

McLafferty,F.W., Fridriksson,E.K., Horn,D.M., Lewis,M.A., and Zubarev,R.A. (1999). Techview: biochemistry. Biomolecule mass spectrometry. Science *284*, 1289-1290.

Meiler,J. and Baker,D. (2005). The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. Journal of Magnetic Resonance *173*, 310-316.

Meiler,J. and Baker,D. (2003). Rapid protein fold determination using unassigned NMR data. Proc. Natl. Acad. Sci. U. S. A *100*, 15404-15409.

Milne,J.S., Mayne,L., Roder,H., Wand,A.J., and Englander,S.W. (1998). Determinants of protein hydrogen exchange studied in equine cytochrome c. Protein Science *7*, 739-745.

Onuchic,J.N., Nymeyer,H., Garcia,A.E., Chahine,J., and Socci,N.D. (2000). The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. Adv. Protein Chem. *53*, 87-152.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B., and Thornton,J.M. (1997). CATH--a hierarchic classification of protein domain structures. Structure *5*, 1093-1108.

Papagrigoriou,E., Gingras,A.R., Barsukov,I.L., Bate,N., Fillingham,I.J., Patel,B., Frank,R., Ziegler,W.H., Roberts,G.C., Critchley,D.R., and Emsley,J. (2004). Activation of a vinculin-binding site in the talin rod involves rearrangement of a five-helix bundle. EMBO J. *23*, 2942-2951.

Parsons,J.T. (2003). Focal adhesion kinase: the first ten years. J. Cell Sci. *116*, 1409-1416.

Perozo,E., Cortes,D.M., and Cuello,L.G. (1998). Three-dimensional architecture and gating mechanism of a K+ channel studied by EPR spectroscopy. Nat. Struct. Biol. *5*, 459-469.

Perrett,S., Clarke,J., Hounslow,A.M., and Fersht,A.R. (1995). Relationship Between Equilibrium Amide Proton-Exchange Behavior and the Folding Pathway of Barnase. Biochemistry *34*, 9288-9298.

Plaxco,K.W., Simons,K.T., and Baker,D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. Journal of Molecular Biology *277*, 985-994.

Pokutta,S. and Weis,W.I. (2002). The cytoplasmic face of cell contact sites. Curr. Opin. Struct. Biol. *12*, 255-262.

Prutzman,K.C., Gao,G.H., King,M.L., Iyer,V.V., Mueller,G.A., Schaller,M.D., and Campbell,S.L. (2004). The focal adhesion targeting domain of focal adhesion kinase contains a hinge region that modulates tyrosine 926 phosphorylation. Structure *12*, 881-891.

Qin,J. and Chait,B.T. (1996). Matrix-assisted laser desorption ion trap mass spectrometry: efficient isolation and effective fragmentation of peptide ions. Anal. Chem. *68*, 2108-2112.

Rahuel,J., Gay,B., Erdmann,D., Strauss,A., GarciaEcheverria,C., Furet,P., Caravatti,G., Fretz,H., Schoepfer,J., and Grutter,M.G. (1996). Structural basis for specificity of GRB2-SH2 revealed by a novel ligand binding mode. Nature Structural Biology *3*, 586-589.

Rapaport DC (1997). The art of molecular dynamics simulation. (Cambridge: Cambridge University Press).

Rapaport,D.C. (2004). The Art of Molecular Dynamics Simulation. (New York: Cambridge University Press).

Reva,B.A., Finkelstein,A.V., and Skolnick,J. (1998). What is the probability of a chance prediction of a protein structure with an rmsd of 6 angstrom? Folding & Design *3*, 141-147.

Ridley,A.J., Schwartz,M.A., Burridge,K., Firtel,R.A., Ginsberg,M.H., Borisy,G., Parsons,J.T., and Horwitz,A.R. (2003). Cell migration: Integrating signals from front to back. Science *302*, 1704-1709.

Sanchez,R. and Sali,A. (1997). Advances in comparative protein-structure modelling. Curr. Opin. Struct. Biol. *7*, 206-214.

Schaller,M.D. (2001). Biochemical signals and biological responses elicited by the focal adhesion kinase. Biochim. Biophys. Acta *1540*, 1-21.

Schlaepfer,D.D. and Hunter,T. (1997). Focal adhesion kinase overexpression enhances Ras-dependent integrin signaling to ERK2/mitogen-activated protein kinase through interactions with and activation of c-Src. Journal of Biological Chemistry *272*, 13189-13195.

Skolnick,J., Kolinski,A., and Ortiz,A.R. (1997). MONSSTER: A method for folding globular proteins with a small number of distance restraints. Journal of Molecular Biology *265*, 217-241.

Smith,A.V. and Hall,C.K. (2001). Protein refolding versus aggregation: Computer simulations on an intermediate-resolution protein model. Journal of Molecular Biology *312*, 187-202.

Smith,S.W., Hall,C.K., and Freeman,B.D. (1997). Molecular dynamics for polymeric fluids using discontinuous potentials. Journal of Computational Physics *134*, 16-30.

Smithbrown,M.J., Kominos,D., and Levy,R.M. (1993). Global Folding of Proteins Using A Limited Number of Distance Constraints. Protein Engineering *6*, 605-614.

Stoesser,G., Tuli,M.A., Lopez,R., and Sterk,P. (1999). The EMBL Nucleotide Sequence Database. Nucleic Acids Research *27*, 18-24.

Subauste,M.C., Pertz,O., Adamson,E.D., Turner,C.E., Junger,S., and Hahn,K.M. (2004). Vinculin modulation of paxillin-FAK interactions regulates ERK to control survival and motility. Journal of Cell Biology *165*, 371-381.

Swaney,J.B. (1986). Use of cross-linking reagents to study lipoprotein structure. Methods Enzymol. *128*, 613-626.

Tachibana,K., Sato,T., Davirro,N., and Morimoto,C. (1995). Direct Association of Pp125(Fak) with Paxillin, the Focal Adhesion-Targeting Mechanism of Pp125(Fak). Journal of Experimental Medicine *182*, 1089-1099.

Tobi,D. and Bahar,I. (2005). Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc. Natl. Acad. Sci. U. S. A *102*, 18908-18913.

Vendruscolo,M., Dokholyan,N.V., Paci,E., and Karplus,M. (2002). Small-world view of the amino acids that play a key role in protein folding. Phys. Rev. E Stat. Nonlin. Soft. Matter Phys. *65*, 061910.

Vendruscolo,M., Paci,E., Dobson,C.M., and Karplus,M. (2003). Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. J. Am. Chem. Soc. *125*, 15686-15687.

Volberg,T., Geiger,B., Kam,Z., Pankov,R., Simcha,I., Sabanay,H., Coll,J.L., Adamson,E., and Ben Ze'ev,A. (1995). Focal adhesion formation by F9 embryonal carcinoma cells after vinculin gene disruption. J. Cell Sci. *108*, 2253-2260.

Wand,A.J. and Englander,S.W. (1996). Protein complexes studied by NMR spectroscopy. Curr. Opin. Biotechnol. *7*, 403-408.

Wang,R. and Chait,B.T. (1994). High-accuracy mass measurement as a tool for studying proteins. Curr. Opin. Biotechnol. *5*, 77-84.

Wasserman,S. and Faust,K. (1994). *Social network analysis : methods and applications* . (Cambridge: Cambridge University Press).

Watts,D.J. and Strogatz,S.H. (1998). Collective dynamics of 'small-world' networks. Nature *393*, 440-442.

Weekes,J., Barry,S.T., and Critchley,D.R. (1996). Acidic phospholipids inhibit the intramolecular association between the N- and C-terminal regions of vinculin, exposing actin-binding and protein kinase C phosphorylation sites. Biochemical Journal *314*, 827-832.

Xu,W., Baribault,H., and Adamson,E.D. (1998). Vinculin knockout results in heart and brain defects during embryonic development. Development *125*, 327-337.

Young,M.M., Tang,N., Hempel,J.C., Oshiro,C.M., Taylor,E.W., Kuntz,I.D., Gibson,B.W., and Dollinger,G. (2000). High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proc. Natl. Acad. Sci. U. S. A *97*, 5802-5806.

Zamir,E. and Geiger,B. (2001). Molecular complexity and dynamics of cell-matrix adhesions. J. Cell Sci. *114*, 3583-3590.

Zhou,Y. and Karplus,M. (1999). Folding of a model three-helix bundle protein: a thermodynamic and kinetic analysis. J. Mol. Biol. *293*, 917-951.

Zhou,Y.Q. and Karplus,M. (1996). Exact results for the effect of bond flexibility on the structure and the collapse transition of isolated square-well trimers. Molecular Physics *89*, 1707-1717.

Zhou,Y.Q., Karplus,M., Wichert,J.M., and Hall,C.K. (1997). Equilibrium thermodynamics of homopolymers and clusters: Molecular dynamics and Monte Carlo simulations of systems with square-well interactions. Journal of Chemical Physics 10691-10708.

Zhu,H., Bilgin,M., and Snyder,M. (2003). Proteomics. Annual Review of Biochemistry *72*, 783-812.